USING BIG DATA TO MODEL TRAVEL BEHAVIOR: APPLICATIONS TO VEHICLE OWNERSHIP AND WILLINGNESS-TO-PAY FOR TRANSIT ACCESSIBILITY

A Dissertation
Presented to
The Academic Faculty

by

Gregory S. Macfarlane

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2014

USING BIG DATA TO MODEL TRAVEL BEHAVIOR: APPLICATIONS TO VEHICLE OWNERSHIP AND WILLINGNESS-TO-PAY FOR TRANSIT ACCESSIBILITY

Approved by:

Professor Laurie A. Garrow, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Professor Juan Moreno-Cruz
School of Economics
Georgia Institute of Technology

Professor Patrick S. McCarthy
School of Economics
Georgia Institute of Technology

Jeffrey P. Newman, Ph.D.
School of Civil and Environmental Engineering
Georgia Institute of Technology

Professor Patricia L. Mokhtarian
School of Civil and Environmental Engineering
Georgia Institute of Technology

Professor Kari E. Watkins
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: January 7, 2014

To Leslie

ACKNOWLEDGEMENTS

In completing this dissertation I am indebted to many different people for their support, be it personal or technical, or in many cases both. I am particularly grateful to Laurie Garrow, who guided me at every stage of this research. She, from my recruitment to my graduation, showed me talents I did not know I had. She always asks me to give more than I sometimes feel I can, and our work together is better for it.

In the times of frustration that occur in any graduate program, I always felt I had somewhere to turn. Kari Watkins and Juan Moreno-Cruz on many occasions shared their perspectives on academic and family life, helping me remain positive and focused when I felt I might not reach this stage.

The students at Georgia Tech are, I feel, the Institute's greatest asset. My classmates here have become lifelong collaborators and some of my closest friends. Donald Katz made spending half my waking hours in a research lab bearable. James Wong kept me humble and reminds me that people will need to use the tools my research creates. Candace Brakewood demonstrates better than anyone how to be involved in many things at once, and how to excel at them all. Thomas Wall shows me how to maintain a mellow attitude while facing difficulties. Josephine Kressner inspires me to a higher standard of technical excellence, and encourages me to be a friendlier person than I might be otherwise.

I am who I am because of my family; my parents Roger and Karen Macfarlane taught me to read, encouraged me to love learning and discovery, and showed me the importance of school when I seemed determined to abandon it entirely.

Finally, this work is dedicated to my wife Leslie, who has paid for its creation more than anyone. Enumerating the ways in which she supports me would be impossible.

TABLE OF CONTENTS

DEDICATION . . . . . **iii**

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **viii**

LIST OF FIGURES . . . . . **ix**

SUMMARY . . . . . **x**

I INTRODUCTION **1**

1.1 Background 1

1.1.1 Big Data 4

1.2 Dissertation Structure 7

1.2.1 Three Studies 7

1.3 Objectives and Contributions 9

1.4 References 10

II THE INFLUENCES OF PAST AND PRESENT RESIDENTIAL LOCATIONS ON VEHICLE OWNERSHIP DECISIONS **12**

2.1 Introduction 13

2.2 Literature Review 15

2.3 Data 18

2.3.1 Motor Vehicle Database 18

2.3.2 Targeted Marketing Data 19

2.3.3 Move Histories 20

2.3.4 Census Data 21

2.3.5 Past Exposure 22

2.3.6 Representativeness 23

2.4 Empirical Model 27

2.5 Results 28

2.5.1 Base Model 29

2.5.2 Models that Include Past Exposure Metrics 29

2.6 Interpretation of Current and Prior Built Environment Effects 33

2.7 Limitations and Future Directions 37
2.8 Conclusions 38
2.9 References 39

III DO ATLANTA RESIDENTS VALUE MARTA? SELECTING AN AUTOREGRESSIVE MODEL TO RECOVER WILLINGNESS-TO-PAY 43

3.1 Introduction 43
3.1.1 Home Prices and Transit Accessibility 46
3.2 Methodology 47
3.2.1 Spatial Autoregressive Models 48
3.2.2 Model Selection 51
3.3 Empirical Application 57
3.3.1 Data 57
3.3.2 Model 58
3.3.3 Spatial Weights 60
3.4 Model results 60
3.5 Discussion 66
3.6 Conclusion 68
3.7 References 68

IV TRANSIT INFRASTRUCTURE AND HOME PRICE STABILITY . 72

4.1 Background 72
4.1.1 Literature 75
4.2 Data 77
4.2.1 Price Performance 78
4.3 Spatial Analysis 81
4.3.1 Results 83
4.4 Latent Class Analysis 86
4.5 Interpretations and Future Directions 91
4.6 References 93

V CONCLUSION 96

5.1 Summary of Findings 96
5.2 Directions for Future Research 97

5.3 References 99

APPENDIX A — APPENDIX TO THE VEHICLE OWNERSHIP PAPER 100

APPENDIX B — APPENDICES TO THE SPATIAL AUTOREGRESSION PAPER 102

VITA 108

LIST OF TABLES

1 Descriptive statistics of quasi-continuous variables. 26

2 Descriptive statistics of categorical variables. 27

3 Base vehicle ownership model. 30

4 Models incorporating exposure metrics. 32

5 Comparison of models that include current attributes and/or past exposure metrics. 34

6 Consequences of misspecified hedonic model. 52

7 Descriptive statistics of model variables 59

8 Estimated model parameters and statistics. 62

9 Lagrange multiplier tests for spatial effects. 63

10 Average marginal effect of model variables. 64

11 Comparison to related studies. 67

12 Volatility metrics for example cases. 79

13 Descriptive statistics of model variables 80

14 Effects of covariates on mean home value. 84

15 Total effects of covariates on home performance. 85

16 Concomitant (class membership) model. 89

LIST OF FIGURES

1 Length of residence in three different databases. 25

2 Geographic distribution of observation and neighbors. 61

3 Estimated elasticity for transit proximity under different specifications, with confidence intervals. 63

4 Case-Shiller home price index, seasonally adjusted. 73

5 Illustrative cases of an abstract investment's value performance. 79

6 Observations in spatial models, by average distance of 50 nearest neighbors. 82

7 Value performance for a sample of 40 homes in each latent class. 88

8 Relative spatial density of homes in each latent class. 90

9 Latent class covariate estimates and confidence intervals. 92

10 Distributions of incomes in the ACS and our estimation sample. 103

11 Log-likelihood under different weighting regimens. 106

12 Autocorrelation and distance to rail parameters under differing weighting regimens. 107

SUMMARY

The transportation community is exploring how new "big" databases constructed by companies or public administrative agencies can be used to better understand travelers' behaviors and better predict travelers' responses to various transportation policies. This thesis explores how a large targeted marketing database containing information about individuals socio-demographic characteristics, current residence attributes, and previous residential locations can be used to investigate research questions related to individuals' transportation preferences and the built environment. The first study examines how household vehicle ownership may be shaped by, or inferred from, previous behavior. Results show that individuals who have previously lived in dense ZIP codes or ZIP codes with more non-automobile commuting options are more likely to own fewer vehicles, all else equal. The second study uses autoregressive models that control for spatial dependence, correlation, and endogeneity to investigate whether investments in public transit infrastructure are associated with higher home values. Results show that willingness-to-pay estimates obtained from the general spatial Durbin model are less certain than comparable estimates obtained through ordinary least squares. The final study develops an empirical framework to examine a housing market's resilience to price volatility as a function of transportation accessibility. Two key modeling frameworks are considered. The first uses a spatial autoregressive model to investigate the relationship between a home's value, appreciation, and price stability while controlling for endogenous missing regressors. The second uses a latent class model that considers all these attributes simultaneously, but cannot control for endogeneity.

CHAPTER I

INTRODUCTION

1.1 Background

The majority of metropolitan planning organizations (MPO) in the United States maintain a travel demand model and a land use forecasting model for transportation planning purposes. The central goal of these models is to identify the consequences of transportation investments and provide a tool for comparing alternative infrastructure or policy regimens. What is the expected ridership for a new transit line? What is the expected revenue from a new congestion pricing scheme? Which highway expansion project delivers the most benefit for the least cost? What would be the land use consequences of expanding a transit network? Aside from providing useful information to policy makers, the analyses performed with travel demand and land use models are often required prior to receiving federal funds or environmental approvals.

A travel demand model takes as inputs land use and population characteristics and computes travel volumes and levels of service. A land use model takes as inputs transportation networks and population characteristics and computes land prices and potential development scenarios. The two types of models work together in a system, and both of these larger model types are themselves a system of smaller sub-models of transportation behavior. A non-exhaustive descriptive list of these sub-models is given below:

Residential Location This model predicts where a household will locate as a function of its socioeconomic characteristics, the work or school locations of its members, the characteristics of the available neighborhoods, and other factors.

Vehicle Ownership This model predicts the number and/or type of vehicles a household owns as a function of its socioeconomic characteristics, the availability of other transportation modes for different trip purposes, and other factors. This model is the subject of Chapter 2.

Trip Generation This model predicts the number and types of trips taken by the members of a household based on the household socioeconomic characteristics and other factors. Also, this model predicts how many trips employment and retail firms will attract.

Trip Distribution This model pairs the productions and attractions output from the Trip Generation model based on network characteristics such as travel times and other factors.

Mode Choice This model predicts the transportation mode used for travel between all origin-destination pairs in the region, based on the characteristics of the available modes for the trip and the socioeconomic characteristics of the travelers, among other factors.

Route Assignment This model predicts the route that the trips will take as a function of network characteristics such as trip costs, transit route transfers, and other factors.

Land Value This model predicts the value of land based on transportation network characteristics, accessibility to retail, employment, and educational activities, and other factors. This model is the subject of Chapters 3 and 4.

Land Development This model predicts which parcels will be developed, renovated, or abandoned based on transportation network characteristics, accessibility to retail, employment, and educational activities, land values, and other factors.

It is possible to refine or combine these models; for example, some recent researchers have attempted to predict household location and vehicle ownership within the same model (e.g. Eluru et al., 2010). The trip generation, trip distribution, mode choice, and route assignment models constitute the core of what has been called a "trip-based," or "four-step," modeling system. It is also possible to introduce models that simulate the movement of individuals throughout the day; the resulting modeling system is called an "activity-based" model. The distinction between a trip-based and an activity-based system — summarized by Bhat and Koppelman (1999) — is not essential in the context of this thesis.

Each of these sub-models relies on estimated parameters that characterize the relationship between the relevant socioeconomic, infrastructure, or accessibility characteristics and the modeled outcome; e.g., a set of parameter estimates will define the effect of an increase in income on the increase in the utility of owning multiple vehicles rather than one. A full transportation modeling system will have hundreds or thousands of potential parameters. Determining which parameters are meaningful and obtaining precise estimates for each of them is a major focus of travel behavior research. The preferred[1] way to estimate the parameters is to collect data from the population that the model seeks to describe, and determine the parameter estimates that provide the best statistical fit and behavioral interpretations. These estimates are then calibrated against observed reality (actual freeway volumes and transit ridership, for instance) to ensure forecasting accuracy.

Collecting a sample of data from the population of interest is itself is a major element of transportation research, but many existing data collection methods are either expensive or unsatisfactory in one or more ways. The primary source of data for regional travel demand models is household travel surveys. In these surveys, each member in a sampled household will record detailed descriptions of his or her trips over a period of one to several days and will also supply demographic information and answers to other questions the researchers are interested in. Typically respondents record their trips in a written diary, but they may also recall the trips during an interview or log them with a researcher-supplied tracking device (for instance, Ogle et al. (2005) placed a GPS unit in the vehicles of survey respondents). These surveys provide valuable information, but at a cost that is becoming prohibitively high for some MPOs; for example, in 2011 the Atlanta Regional Commission conducted region-wide household travel survey that cost approximately $2 million, or $200 per response (Rousseau, 2010). Household travel surveys may also exhibit diminishing returns, in that increasing the detail of the survey may weaken the quality of the received data through respondent survey fatigue. Finally, the regional nature of these surveys means that if city A is going to implement the modeling advances of city B, it may need to conduct its own

[1] When it is not possible to estimate a parameter, the analyst sometimes uses her judgment to select an estimate from the literature or another region's model.

survey to estimate the necessary parameters.

The federal government runs several national programs to collect and disseminate data that can be used in transportation behavior models. These programs include the Census Bureau's American Community Survey (ACS), the Federal Highway Administration's National Household Travel Survey (NHTS), the Department of Housing and Urban Development's American Housing Survey (AHS), and many others. The resulting data products are free to the public but cannot provide the detail of household surveys. The ACS, for instance, collects data on transportation mode and trip distance, but only for work-related trips. Another weakness is that the operating agencies are concerned with personal information disclosure, and their non-disclosure methodologies make the data difficult to use for transportation modeling. Disaggregate ACS responses are available through the public use microsample (PUMS) data products, but the geographic resolution is censored to an area containing at least 100,000 residents; modeling the origin and destination of most urban trips is therefore impossible. Finer resolutions are available for aggregate data, but then individual travel behavior is not available. These restrictions make Census and other federal products more useful in developing sampling weights and calibrating model outputs. These are essential tasks, but are secondary to the problem of understanding how individuals' travel behavior is influenced by various factors, and how these individuals may respond to policy changes.

1.1.1 Big Data

A distinguishing feature of contemporary society is the pervasive digitization of personal information and recording of individual behavior in various electronic databases. These databases may be created as a by-product of an unrelated commercial service, to streamline an administrative responsibility, or to expressly provide socioeconomic information to other firms (illustrative examples are given below). Collectively, these databases contain much of the information needed to develop travel behavior or land use models. These databases may be considered a form of "Big Data." Though this term is as widely used as it is loosely defined, a popular definition by the consultancy Gartner (2013) classifies big data as

> ... high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Compared to a standard regional household travel survey, modeling data derived from these databases meet these criteria. Rather than a sample of observations, the databases contain a record of all relevant individuals and transactions (*volume*). Whereas a survey is a static record of the sample, the databases are constantly updated with new records or observations (*velocity*). Finally, analysis of survey data is limited to the questions asked on the survey questionnaires; a big data analysis methodology is instead limited by the researcher's ability to identify relevant information in existing and future databases (*variety*). Rather than collecting data themselves, transportation modelers might access databases maintained by third-party data providers. Further, a researcher that joins information from multiple databases together may study questions that are not addressable with any single database alone. Three classes of database may be of particular relevance to transportation behavior studies: public administrative databases, targeted marketing databases, and device location databases.

Targeted Marketing Records In the process of assembling household financial credit reports, several companies collect personal financial and demographic information that is needed to establish the credit worthiness for virtually all adults in the U.S. Much of this information may be of interest to other businesses that want to target their services to specific customer segments. Some credit reporting firms and several targeted marketing firms provide databases of individuals' characteristics. Information commonly contained in these databases includes:

- Socioeconomic data such as income, occupation, and education level.
- Household structure, including number and ages of children.
- Home address, housing type, and home value.

Other information that may be available through some firms includes spending habits and the previous addresses of household members. Many of these items are collected directly from an individual's financial records, but others come from self-reported information such as loan applications, address changes, and product warranty registrations.

Public Administrative Databases Much of the information used to develop travel demand or land use models is already collected by public agencies through the course of their administrative responsibilities. As an example, counties or cities assess the value of homes each year in order to levy property taxes. This information is typically considered public record and has long been used to study the relationship between the housing market and infrastructure investment (e.g., Iacono and Levinson, 2011; Bitter et al., 2006); the increasing digitization of this information provides new opportunities for more comprehensive and comparative studies.

Motor vehicle departments similarly maintain registration databases recording when vehicles were bought and sold, and to whom. This information could be used to study vehicle ownership decisions. In regions required to mitigate vehicle emissions under the Clean Air Act, the results of emissions tests are also typically recorded in these or related databases. These databases can therefore be used to study emissions testing policies (e.g., Washburn et al., 2001; Choo et al., 2007). When joined with demographic information (such as TM records), they can be used to examine the social implications of such policies; see Binder et al. (2014) as an example of such a study.

Device Location Databases Many individuals carry electronic devices that record their location at several points and times during the day. Transit agencies have been using data collected from electronic fare payment cards (smartcards) to observe the trip making characteristics of their passengers, such as origin, time of departure, destination, and trip duration (Seaborn et al., 2009). In-vehicle GPS units give similar information on automobile trips, with the added ability to observe route choice, travel time, travel time reliability, and other factors (Byon et al., 2006). Cellular telephone carriers observe the location of phones hundreds of times per day as the phones check in with physical towers. This information records

points along an individual's route and can be used to derive the origin, destination, route, and other associated trip information for motorized and non-motorized trips irrespective of mode (Smith et al., 2005). Developing a methodology for joining this trip information to targeted marketing records could in principle provide much of the information usually collected in regional household travel surveys.[2]

1.2 Dissertation Structure

The dissertation contains a total of five chapters and two appendices. Chapters 2 through 4 contain three studies focusing on the relationship between transportation infrastructure and observed individual or market preferences. These studies are presented in a journal format: each study contains its own motivation, econometric approach, findings, and references to the existing literature.

1.2.1 Three Studies

The first study, presented in Chapter 2, considers the relationship between residential land use and vehicle ownership. This study specifically tests the hypothesis that *prior* exposure to high density land uses or alternative transportation modes affects current vehicle ownership preferences. The study relies on a dataset compiled from household credit records that reveals the address histories for a large sample of Atlanta residents, combined with the state's vehicle registration database. The data show that households that have previously lived in areas where vehicle ownership is reduced currently own fewer vehicles, all else equal. This finding indicates that people may learn their preferences through experience, though the effects of self-selection of residents into neighborhoods means the magnitude of this effect is modest.

The second study, presented in Chapter 3, considers the willingness of households to pay

[2]Though this dissertation does not utilize data of this type, the spatial location of daily activities is the motivating purpose for transportation, and even a potential means of recovering this information from a third party deserves comment.

for access to public transportation infrastructure through their home price. This willingness-to-pay is an important parameter of land use models and infrastructure investment strategies, but econometric complications resulting from spatial effects make estimating this parameter difficult. To assess the importance of correct specification, this study applies a database of targeted marketing records that reveal individual homeowner characteristics, including the value of a home. The assessment shows a classical selection framework built on testing for spatial effects may select a non-optimal model specification, substantially affecting the estimate of willingness to pay. Future researchers should therefore adopt a framework that relies on testing reductions to the general spatial Durbin model. The spatial Durbin model estimates reveal a strong and significant willingness to pay for proximity to public transit infrastructure in the Atlanta home market.

The final study, presented in Chapter 4, investigates the resiliency of a home's price as a function of its transportation network accessibility. The U.S. home market in the 21st century to the present can be succinctly described by a large increase in prices from 2000 to 2007-2008 and a sudden crash from which the market has not yet recovered. This national story is not homogeneous, however. Average home prices in some cities remained effectively unchanged, and in some cities prices collapsed without first experiencing a boom. Even within a single market, the price outcome in some neighborhoods was more favorable than in others. Could the built environment, and particularly public transit infrastructure, play a role in determining this outcome? If access to public transportation is a good that homeowners are willing to pay for, then neighborhoods with good access may behave differently from the rest of the market. The third and final study presents an empirical framework for analyzing this question. Spatial Durbin models of value, price volatility, price growth, and growth volatility show the effect of transit proximity on each dimension separately while simultaneously controlling for spatially endogenous omitted variables. Latent class models show the relationship between transit proximity and all dimensions simultaneously. These models show that homes near MARTA stations are more likely to have accumulated value over their initial price in the period from 2002-2012, and homes further away are more likely to have collapsed in value.

The dissertation concludes in Chapter 5 with a summary of major findings and directions for future research. Appendix A presents a sensitivity analysis on the assumptions made in Chapter 2; Appendix B presents an analysis discussing the selection of a spatial weights matrix for the spatial hedonic models estimated in Chapter 3.

1.3 Objectives and Contributions

Collectively, the dissertation has two primary objectives. The first objective is to explore the potential and the feasibility of using certain third-party big data products in different types of travel behavior or land use models. This is an expressed research need, with the Transportation Research Board of the National Academies calling for the

> [development] of methods for enhancing the quality of activity and travel data collected through new technologies, making sense of it, and using it to enhance our travel demand forecasting models.

The second objective is to explore the additional questions that can be addressed with third-party big data products that might not be easily explored with traditional household travel surveys. Credit records contain address histories of households, a variable that can provide insight into experience and preferences but which may be difficult to capture in surveys. Further, the precision of parameter estimates is a function of the number of observations; the sheer size of big data products may allow researchers to reject null hypotheses that could not be rejected with smaller surveys. The number of observations also aids in estimating models on segments or subpopulations, further increasing the precision of the model estimates.

The contributions of the dissertation are symmetrical to the objectives. In terms of the models that can be estimated, Chapter 2 presents a vehicle ownership model estimated on targeted marketing records joined to a state vehicle registration database. Chapters 3 and 4 present land value models estimated on targeted marketing records and a county assessors database, respectively. The vehicle ownership model closely replicates comparable models found in the literature (e.g., Ryan and Han, 1999; Dieleman et al., 2002), confirming the prior results and illustrating the usefulness of the targeted marketing data. The land value

models also show a positive willingness to pay for transit accessibility, confirming the results of several previous studies (a review and meta-analysis is given by Debrezion et al. (2007)).

The big data resources used in the studies have attributes that permit the thesis to reach beyond mere replication, however. The prior addresses of the individuals in Chapter 2 unlocked a relationship between an individual's prior experience and their current vehicle ownership behavior that had been theorized but not comprehensively addressed (e.g., Weinberger and Goetzke, 2010). The disaggregate demographic data used in Chapter 3 allows for a fine resolution in the socioeconomic variables used as controls, highlighting the distinction between direct and indirect effects. Finally, the digitization of the assessments of every home in Fulton County permits the models in Chapter 4 to be quickly and conveniently estimated; the number of observations further aids the ability to infer the significance of relatively small parameter estimates.

1.4 References

Bhat, C.R., Koppelman, F.S., 1999. Activity-based modeling of travel demand, in: Hall, R.W. (Ed.), Handbook of Transportation Science. Springer US, pp. 35–61.

Binder, S., Macfarlane, G.S., Garrow, L.a., Bierlaire, M., 2014. Associations among household characteristics, vehicle characteristics and emissions failures: An application of targeted marketing data. *Transportation Research Part A: Policy and Practice* 59, 122–133.

Bitter, C., Mulligan, G.F., Dallerba, S., 2006. Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems* 9, 7–27.

Byon, Y., Shalaby, A., Abdulhai, B., 2006. Travel Time Collection and Traffic Monitoring Via GPS Technologies, in: 2006 IEEE Intelligent Transportation Systems Conference, IEEE. pp. 677–682. URL: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1706820`.

Choo, S., Shafizadeh, K., Niemeier, D., 2007. The development of a prescreening model to identify failed and gross polluting vehicles. *Transportation Research Part D: Transport and Environment* 12, 208–218.

Debrezion, G., Pels, E., Rietveld, P., 2007. The impact of railway stations on residential and commercial property value: a meta-analysis. *The Journal of Real Estate Finance and Economics* 35, 161–180.

Dieleman, F.M., Dijst, M., Burghouwt, G., 2002. Urban form and travel behaviour: micro-level household attributes and residential context. *Urban Studies* 39, 507 – 527.

Eluru, N., Bhat, C.R., Pendyala, R.M., Konduri, K.C., 2010. A joint flexible econometric model system of household residential location and vehicle fleet composition/usage choices. *Transportation* 37, 603–626.

Gartner, 2013. Big Data. URL: http://www.gartner.com/it-glossary/big-data/. accessed November 25, 2013.

Iacono, M., Levinson, D., 2011. Location, regional accessibility, and price effects: Evidence from home sales in Hennepin County, Minnesota. *Transportation Research Record* 2245, 87–94.

Ogle, J., Guensler, R., Elango, V.V., 2005. Georgia's commute Atlanta value pricing program: Recruitment methods and travel diary response rates. *Transportation Research Record* 1931, 28–37.

Rousseau, G., 2010. Atlanta Household Travel Survey. Presented to the ARC TCC Meeting, October 8, 2010.

Ryan, J.M., Han, G., 1999. Vehicle-ownership model using family structure and accessibility: Application to Honolulu, Hawaii. *Transportation Research Record* 1676, 1–10.

Seaborn, C., Attanucci, J., Wilson, N.H.M., 2009. Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record* 2121, 55–62.

Smith, C.W., Wilkinson, C., Carlson, K., Wright, M.P., Sangal, R., 2005. System and method for providing traffic information using operational data of a wireless network. United States Patent # 6,842,620 B2.

Washburn, S., Seet, J., Mannering, F., 2001. Statistical modeling of vehicle emissions from inspection/maintenance testing data: An exploratory analysis. *Transportation Research Part D: Transport and Environment* 6, 21–36.

Weinberger, R., Goetzke, F., 2010. Unpacking preference: How previous experience affects auto ownership in the United States. *Urban Studies* 47, 2111–2128.

CHAPTER II

THE INFLUENCES OF PAST AND PRESENT RESIDENTIAL LOCATIONS ON VEHICLE OWNERSHIP DECISIONS

Gregory S. Macfarlane, Laurie A. Garrow, Patricia L. Mokhtarian

Working Paper, 2014

Chapter Abstract

The correlation between certain land use types and vehicle ownership patterns creates an important tool for planners seeking to change transportation behavior if the relationship is causal, but there are alternative explanations. People may primarily select to live in neighborhoods that facilitate their vehicle ownership preferences, or they may retain behaviors that they have learned in the past, irrespective of current situation. This study considers these alternative explanations by measuring the influence of past and present residential locations on current vehicle ownership. We use a dataset from a credit reporting firm that contains up to nine previous residential ZIP codes for individuals currently living in the 13-county Atlanta, Georgia, metropolitan area. Results show that past experience is significant, but of marginal influence once controlling for the current location. Conversely, controlling for past locations, the current location still has a much stronger contribution in explaining vehicle ownership. From a practical perspective, our results suggest that models that include "only" current neighborhood attributes (in addition to standard socioeconomic variables) can accurately forecast vehicle ownership decisions. However, our results also show that models that include both current and past neighborhood attributes, along with other variables, can provide a more nuanced understanding of causal influences on vehicle ownership decisions. In turn, this can help policy-makers better design and target strategies for reducing vehicle ownership among particular groups of individuals.

2.1 Introduction

Society's dependence on private vehicles creates several negative externalities. From an economic perspective, traffic congestion cost the U.S. economy $121 billion in lost wage productivity in 2011 (Schrank et al., 2012). Economic externalities from vehicle dependence may be even more pronounced among certain demographic groups; for instance, the reduced ability of low-income households to obtain vehicles is often viewed as a contributing factor to low economic mobility (Leonhardt, 2013; Matas et al., 2009). In addition, urban air pollution is largely due to transportation-related emissions and can contribute to global climate change (Chapman, 2007), respiratory ailments (Buckeridge et al., 2002), and other negative externalities.

Previous studies have explored relationships among vehicle ownership and the built environment (e.g. Ewing and Cervero, 2001; Dieleman et al., 2002; Giuliano and Dargay, 2006; Van Acker and Witlox, 2010). These studies hypothesize that individuals who live in higher-density or more mixed-use areas can access a larger number of activity locations by walking, biking, or taking public transit, thereby reducing the need for one (or more) vehicles. Policies designed to increase densities or land use mix are often viewed as mechanisms for reducing vehicle ownership and/or vehicle usage, which in turn would help reduce emissions (Kenworthy and Laube, 1996; Norman et al., 2006; Stone, 2009), potentially alleviate congestion, and improve transportation equity (Sanchez et al., 2003).

To isolate the autonomous influence of the built environment on vehicle ownership decisions, it is important to control for other possible causal influences. On one hand, self-selection could explain part of the observed correlation between the built environment and travel behavior; that is, individuals who prefer to own fewer vehicles may choose to live in denser or more mixed neighborhoods *so that* they can own fewer vehicles. Density in this situation facilitates, rather than causes, a particular behavior. If this is true, then incentivizing or requiring density through zoning or tax policies may not change vehicle ownership in a meaningful way, *unless preferences also change*. On the other hand however, individuals' preferences for vehicle ownership may, in fact, evolve over time as they are

exposed to more dense and mixed neighborhoods and learn about non-vehicle transportation options. If this is true, then building denser or more mixed-use developments could eventually lead to lower levels of vehicle ownership, although the short-term effectiveness of using density as a planning tool for positive environmental and economic changes could be diminished.

In this study, we use multinomial logit models to predict the (non-zero) number of vehicles owned by a household as a function of the head of the household's socioeconomic characteristics, exposure to residential densities (defined as housing units per square kilometer of the current and previous residential locations) and exposure to non-vehicle alternatives (as revealed through the commuting mode shares of the current and previous residential locations). Results help establish a secondary association between the built environment and vehicle ownership: all else equal, a household that has experienced higher densities and higher non-vehicle commuting mode shares *in the past* owns fewer vehicles *in the present*. Overall, the influence of past exposure on vehicle ownership decisions (which can be interpreted as the joint effect of learned preferences and self-selection) is smaller than the influence of current exposure. From a practical perspective, these findings suggest that vehicle ownership models that include just current built environment characteristics (in addition to standard socioeconomic variables) should be able to accurately forecast vehicle ownership. Conversely, researchers who are able to construct an analysis database of socioeconomic variables, attributes of both the current and past built environments can provide a more nuanced understanding of potentially causal influences on vehicle ownership decisions.

The paper is organized into several sections. Section 2.2 describes how our study contributes to the literature. Section 2.3 provides an overview of the analysis database and data processing assumptions. Sections 2.4 and 2.5 follow, presenting the econometric methodology and results, respectively. Section 2.6 interprets the empirical results in the context of the learned preference and self-selection theories, and Section 2.7 presents a discussion of study limitations and directions for future research. The paper concludes with a summary of key findings and implications for practice.

2.2 Literature Review

Many models view vehicle ownership as a strictly utilitarian phenomenon. In this perspective, a given household has a need for vehicles established by the size of the household (or its number of workers) and the availability of vehicle alternatives. The household acquires the necessary vehicles subject to income constraints (as an example, see Potoglou and Susilo, 2008). In contrast, a growing body of literature shows that vehicle owners derive considerable affective utility from the type of vehicle they own, predicated on their attitudes and preferences. Some people have a propensity to own vehicles that are faster or more stylish than strictly necessary (Lois and López-Sáez, 2009). Others choose vehicles that signal environmental-political preferences (Sexton and Sexton, 2011). More generally, it has been shown that people tend to own vehicles that are similar to those driven by their neighbors (Adjemian et al., 2010).

If owning a particular vehicle is a reflection of attitudes and preferences, then it follows that not using or even owning a vehicle reflects a preference as well. This was highlighted by Fujii and Kitamura (2003), who showed that providing a free transit pass to habitual vehicle commuters permanently changed the behavior of many. Their finding indicated the study participants' prior vehicle usage was at least partly attitudinal (at least for those who changed), rather than purely utilitarian. Having experienced a transportation alternative that was acceptable and perhaps even superior from the utilitarian perspective, these individuals adjusted their attitudes and preferences. A profound implication of this study is that vehicle ownership and usage (or vehicle independence) is — to some extent — learned.

If learned preferences are a mechanism by which vehicle ownership decisions are made, then characteristics of the built environment at an individual's previous residential location(s) should help explain current vehicle ownership decisions. This is the modeling approach taken by Weinberger and Goetzke (2010), who examined the relationship between implied exposure to vehicle independence and current vehicle ownership using national data from the 2000 Census "long" form survey (U.S. Census Bureau, 2000a). The 2000 Census asked a sample of respondents "Did this person live in this house or apartment 5 years ago (on April 1, 1995)?" If the respondent answered "no," a follow-up question obtained the

address of the location where the individual on April 1, 1995. Weinberger and Goetzke showed that people who reported prior addresses in San Francisco, Chicago, Philadelphia, Boston, New York, and Washington, D.C. owned fewer cars than other households, all else equal. The authors attributed the influence of prior exposure on vehicle ownership to a learned preference mechanism.

An analogous explanation to learned preferences, and one that has received increased attention in the literature, relates to self-selection. Under a self-selection paradigm, individuals are assumed to be pre-disposed to live in a certain type of built environment: some individuals prefer to live in high density urban areas whereas other individuals prefer living in low density suburban areas. See Cao et al. (2009) and Mokhtarian and Cao (2008) for review articles in this area. A recent study that focuses on self-selection and vehicle ownership is given by Cao and Cao (2013), and examples of studies that examine the effect of self-selection on other transportation behaviors include Handy et al. (2006) and Pinjari et al. (2009). Research in this area seeks to determine: (1) whether the built environment has a distinct influence on travel behavior after self-selection is accounted for and, if so, (2) the strength of the autonomous influence of the built environment relative to the influence of self-selection.

There are two main approaches to measuring or controlling for the self-selection problem. The first relies on econometric techniques that build self-selection endogeneity directly into mathematical relationships; these techniques may include structural equations models (Bagley and Mokhtarian, 2002; Cao et al., 2007b) or discrete choice models that jointly estimate residential choice and vehicle ownership (e.g., Eluru et al., 2010; Roorda et al., 2009). The second approach is to include individuals' preferences or experiences with residential urban forms directly in the model as one or more exogenous variables. Preferences can be measured through direct inquiry or can be inferred by observing a household's previous residences. For example, Cao et al. (2007a) surveyed 547 households that had recently (in the previous year) moved into a group of Northern California neighborhoods representing either traditional or suburban land use characteristics. These characteristics included indicators such as the age and style of homes, street connectivity, and distance to various commercial

establishments. The survey asked the households about their current and previous vehicle ownership levels as well as their attitudes towards travel behaviors (e.g., "I need a car to do many of the things I like to do") and neighborhood design (e.g., "I prefer shopping areas within walking distance"). The authors showed that these attitudes were more predictive of household vehicle ownership than were objective measurements of the neighborhoods. Additionally, the households in the survey tended to relocate to either traditional or suburban areas in a pattern consistent with both their previous land use and their expressed attitudes. Nevertheless, the authors allow that preferences might change with experience, stating "it is possible that the built environment also plays an additional indirect role by influencing these attitudes over time" (Cao et al., 2007a, p.846).

Note that in both the self-selection as well as the learned preference paradigm, exposure measures that describe characteristics of the built environment at past residential locations may help to infer the relative magnitude of the impact of the current built environment on vehicle ownership decisions. A household's previous exposure to density, for example, may either indicate what its historical preferences have been or may signify experiences that have shaped those preferences. However, with exposure measures alone, it is not possible to distinguish how much of the influence of past residential locations is due to a learned preference mechanism and how much is due to a self-selection mechanism.

In this paper, we build on the two previous papers we are aware of (Weinberger and Goetzke, 2010; Cao et al., 2007a) that have used previous addresses to investigate the influence of past and present residential locations on current vehicle ownership decisions. In contrast to these previous studies, our dataset contains multiple prior residential locations as well as the length of time spent at each residential location. This enables us to derive a wide set of exposure metrics and test the robustness of results to different exposure metrics. In this study, we interpret historical exposure as confounding elements of self-selection and learned preferences, that is, we are agnostic as to the dominant operating theory.

2.3 Data

Data that would provide researchers with a better ability to infer the autonomous influence of the built environment on vehicle ownership decisions would ideally satisfy at least the following two criteria: (1) a history of several previous addresses, including the dates of relocation; and, (2) a nationwide scope of previous addresses given at a small geographic resolution. Our analysis database satisfies these two criteria. The database is compiled from four primary sources: vehicle information is obtained from the state's registration database, socioeconomic information is obtained from a targeted marketing (TM) firm, residential move histories are compiled from a credit reporting firm, and residential densities and commuting mode shares are calculated from Census data. Our final analysis dataset contains 227,830 households, constituting a 12% sample of the 13-county metropolitan Atlanta region. Definitions for each of the variables used in the study, as well as descriptive statistics for these variables, are provided in Tables 1 and 2. The majority of the variables shown in the tables have a straightforward interpretation. This section provides definitions for variables that merit additional discussion and provides an overview of key assumptions used to compile the analysis database that may influence the representativeness of our analysis database and/or the interpretation of results.

2.3.1 Motor Vehicle Database

The number of vehicles owned by a household was obtained from the Georgia Department of Revenue's Motor Vehicle Division (called the DMV), which maintains records for all vehicles registered with the state. For the purposes of this study, we consider as vehicles only passenger cars and light-duty trucks (not, for instance, cargo vans or motorcycles). We drew a simple random sample of vehicles registered to addresses in the 13-county metropolitan Atlanta area from this database, removed duplicate registration addresses, and simply sampled replacements. We then appended to each sampled address the full list of registered vehicles associated with that address as of December, 2010. Home addresses were standardized prior to merging the databases, however approximately 1.5% of the addresses had to be excluded as it appeared that apartment numbers were omitted for some households.

During the merge process, this resulted in a large number of vehicles being associated with multi-unit buildings. Although it was not possible to determine exactly which addresses represented unique households (versus a multi-unit building), we assumed those addresses with five or fewer vehicles represented unique households.

2.3.2 Targeted Marketing Data

Target marketing (TM) firms compile information about individuals from a variety of sources (e.g., public records, product registration cards, credit card transactions). These data are often sold to advertising companies to customize marketing campaigns to potential customers. These TM databases contain the majority of household and individual demographic fields that are used in travel demand forecasting applications. TM data has been used in several prior travel demand studies (e.g., see Binder et al., 2014; Kressner and Garrow, 2012). We appended the TM records for the addresses sampled from the motor vehicle database; addresses that could not be matched to a consumer (for instance, vehicles registered to a business) were returned with no additional information appended.

The TM database provides socio-economic information about the number of adults in the household, the number of children in the household, the household annual income and ethnicity, the age of the head of household, and housing tenure. As an explanatory variable in the vehicle ownership model, we include the number of adults in the household in relation to the number of vehicles that could be potentially owned by the household. Specifically, for a potentially-chosen number of vehicles j in household i, we define $Insufficiency_{ij}$ as

$$\max(0, Adults_i - j) \tag{1}$$

For example, if there are two adults in the household, the insufficiency variable would take a value of one for the "one vehicle" alternative ($j = 1$), zero for the "two vehicles" alternative, and zero for the "three or more vehicles" alternative. In this sense, the insufficiency variable represents a measure of competition for limited vehicle resources in the household. Households with no income constraints or intervening preferences will attempt to minimize this value (or choose the number of vehicles to at least equal the number of adults in the household).

For the TM database used in this study the variable indicating the number of children is ambiguous in that a "0" value may indicate either a known zero count or an unknown value.[1] An analysis against Census data (U.S. Census Bureau, 2012a) revealed that 38% of households in the Atlanta region have children, whereas in our data only 28% do. We examined the robustness of model results to different imputation methods for this variable as part of the analysis.

2.3.3 Move Histories

The TM firm who provided data for this analysis has a close relationship to a credit reporting firm. The credit reporting firm maintains a database of current and previous addresses at the ZIP code level for (in theory) up to four adults living in the household. A maximum of nine previous ZIP codes for each adult is available. The date (month and year) that each address changed is also available. For households with more than one adult listed in the TM records, only 40% contained a second address history. Using this second history would potentially allow us to study mismatched households: for instance, a household where one partner has lived in dense urban areas and the other has not. However, using the address history for the second adult in the household would require us to make strong assumptions about household formation and dynamics, so we restricted our analysis to the move histories for the heads of households.

We made several assumptions to construct coherent address paths from the source data. The credit reporting firm noted that the ZIP codes and associated move-in dates may not be in sequential order for all households. Of the records in our dataset, 34% required reordering. Also, by comparing the current ZIP codes in the vehicle registration and credit reporting databases, we noted inconsistencies in how the credit company recorded the "most recent" address. Although the majority of "most recent" ZIP codes matched 27% did not. We assumed this implied that the credit reporting database was missing the most recent move, i.e., that the ZIP code obtained from the TM database was accurate. We calculated the move-in date for these households using two assumptions: the first assumption represents

[1]The TM marketing firm recently updated the algorithm it uses to populate the number children field, so this is not expected to be a limitation in subsequent studies.

the "earliest" move-in date and the second assumption represents the "latest" move-in date. The earliest move-in date assumes the move occurred one month after the head of household's last known address change. The latest move-in date assumes the move occurred one month before the data were collected, or on December 1, 2010. We tested the sensitivity of exposure metric calculations to these two assumptions. Finally, some households in the database (1.4%) had consecutive moves in the same month; we forced the length of residence in this case to be one day.

The prior ZIP codes provide comprehensive coverage of the U.S. On average, the heads of households in the database have lived at 2.5 previous addresses. There are 14,785 unique prior ZIP codes in the database, which represents 45% of all ZIP codes in the U.S. ZIP codes from all fifty states, the District of Columbia, and Puerto Rico are represented at least once. The large majority of ZIP codes (89%) are from the southeastern U.S.,[2] specifically Georgia (73%) and the Atlanta metropolitan statistical area (MSA) itself (71%). Excluding prior addresses in the Atlanta MSA, 32% are from other MSAs that are ranked in the top ten by population and 62% are from MSAs that are ranked in the top fifty. The six cities identified by Weinberger and Goetzke (2010) as having high transit accessibility account for 21% of the non-Atlanta ZIP codes.

2.3.4 Census Data

As land use and transportation behavior may independently affect preference development, it is important to have a measurement of each. Our measurement of land use is density, which we define as the count of housing units in each ZIP code divided by the land area (in square kilometers) of the ZIP code. As a measure of the availability or feasibility of non-vehicle transportation modes, we define the "alternative mode share" as the proportion of workers who *do not* drive or carpool to work in each ZIP code. These people may use public transit, walk, bike, etc. We calculate these metrics from Census data. For the current ZIP code and for residences occupied since 2005, we use tables from the American Community

[2]Defined by Census as being the area bounded by and including Texas, Oklahoma, Arkansas, Kentucky, West Virginia, and Maryland.

Survey (U.S. Census Bureau, 2012b, 2011).[3] For residences occupied earlier than 2005, we draw instead from the 2000 Census and its "long" form sample (U.S. Census Bureau, 2000b,c).

2.3.5 Past Exposure

This study tests the hypothesis that a household's previous experience with density or exposure to alternative transportation behavior can be used to predict its current vehicle ownership. We do this by creating a household *past exposure metric*, E_i, for household i. This metric applies some rule to characterize the density or alternative share at all of the previous addresses in which the head of household is known to have resided. As the existing literature provides little insight into how prior experience shapes preferences for vehicle ownership, we develop an array of plausible exposure metrics. Because households may be observed for different lengths of time and/or have a different number of prior addresses, we use metrics that that do not require comparisons across households.

To describe our past exposure metrics, we first need to formalize notation. Given household i that lived at previous ZIP code number $z, z = 1, 2, 3 \ldots$, the density or non-vehicle mode share associated with the previous ZIP code, d_{iz}, is defined for up to nine previous addresses. However, because not all households have that many addresses in the database, we define Z_i as the maximum number of previous addresses available for household i. The corresponding length of residence at ZIP code number z is denoted as t_{iz}. Households with no previous addresses, and therefore no observable past exposure, are excluded from the study.

Duration The duration exposure score is the arithmetic mean of the attribute, weighted by the time spent in the corresponding ZIP code. Longer residences carry more weight, but each additional day has an unknown marginal importance α:

$$E_i = \frac{\sum_{z=1}^{Z_i} d_{iz} * t_{iz}^{\alpha}}{\sum_{z=1}^{Z_i} t_{iz}^{\alpha}}. \tag{2}$$

[3]The 5-year aggregation for 2007-2012 is used as it is the earliest ACS product to provide tables by ZIP Code Tabulation Area (ZCTA).

In this formulation, $\alpha = 1$ represents a constant marginal effect with each day carrying equal weight. For $\alpha > 1$ there is an increasing marginal effect for each day, thus accentuating the relative importance of long residences; this may represent individuals developing an "addiction" to the attribute. For $0 < \alpha < 1$ there is a diminishing marginal effect, increasing the relative importance of shorter residences. Finally, the case where $\alpha = 0$ reduces the exposure score to the simple arithmetic mean, which indicates that the exposure is duration-insensitive and that each prior residence has equal weight.

Decay The decay exposure score assumes that memory decays, or that older addresses affect current behavior less than recent ones. We use a geometric decay function:

$$E_i = \frac{\sum_{z=1}^{Z_i} d_{iz} * t_{iz}/z}{\sum_{Z_i}^{9} t_{iz}/z}. \tag{3}$$

Under this formulation each day spent at the $z = 3$ (third prior) address contributes 1/3 as much to the exposure as each day at the $z = 1$ (first prior) address; similarly, each day spent at the $z = 5$ (fifth prior) address contributes 1/5 as much as each day spent at the $z = 1$ (first prior) address.

Extreme The extreme exposure score assumes that individuals' behavior is most influenced by the highest density or alternative share they have ever experienced, or

$$E_i = \max(d_{iz}), z = 1, \ldots, Z_i. \tag{4}$$

Longest The longest exposure score assumes that individuals' behaviors are simply a function of the place z^* at which they have resided at the *longest*, or that

$$E_i = d_{iz^*} \tag{5}$$

where z^* is the prior residence for $t_{iz^*} = \max_z(t_{iz}), z = 1, \ldots, Z_i$.

2.3.6 Representativeness

After assembling data from the four sources described above, our database contained records for 417,538 households (representing 22% of all Atlanta households). However, for a record

to be useful in our analysis, the variables of interest for each record must be known. In this study, we deleted records with one or more missing fields, as the resulting dataset of 227,830 complete records was still fairly representative of the Atlanta region as a whole, particularly with respect to income. To illustrate this point, in our estimation sample the median income is $62,500 per year and the mean is $80,196; for vehicle-owning respondents to the ACS in the 13-county Atlanta region,[4] the comparable figures are $62,000 and $81,693, respectively. The distributions of the number of household adults and the ages of adults in the household are also similar between the estimation sample and Census data. However, our sample is biased towards homeowners (95% versus 78%) and Whites (75% versus 67%). This bias was also seen in a study by Kressner and Garrow (2013), which compared a "complete" sample of TM data at the block group level with Census data for the 13-county Atlanta region; that is, the over-representation of homeowners and Whites is seen in the "full TM database" as well as in the "reduced analysis dataset" ultimately used for this study.

To assess whether the length of time at the current residence calculated from the credit reporting database was representative, we compared the distribution of this variable to housing tenure fields available in the TM and Census data. The TM data contain a variable that is the length of time the household has lived at its current address. This variable is correlated with (correlation 0.62), but not identical to, the length of time calculated from credit records. Upon further observation, the credit records tend to show a higher proportion of households with shorter tenures at the current residence than the TM data. Figure 1 shows a comparison of the length of residence as recorded in the TM database, the credit records, and as measured in the 2009 American Housing Survey (AHS). The AHS data are for the entire Southeast region (the smallest relevant geography available to the public), but Atlanta residents are likely more mobile than the Southeast region in general. The credit records show a higher proportion of households with shorter residences than either the AHS or TM records. A plausible explanation is that some people move away only temporarily (e.g., college students living on campus during the academic year and

[4]Specifically, the collection of PUMAs that contain the 13 counties as defined by the Environmental Protection Agency for air quality mitigation purposes.

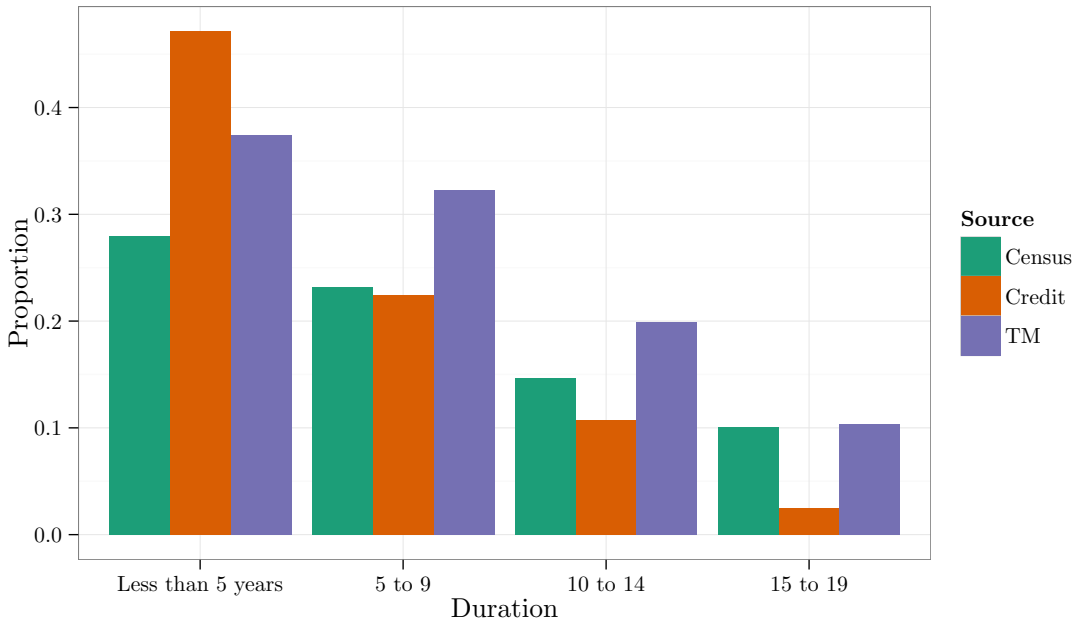

Figure 1: Length of residence in three different databases.

at their parents' home during break); in this case, their credit records may show multiple relocations that may not be self-reported.

Although the sample is not representative of the population in every way, that is less of a concern when the purpose of the sample is to uncover relationships among variables (as it is here) than when it is purely to describe a population (Babbie, 2009; Groves, 1989, Chap. 1). For example, if we were using the sample to estimate the true share of various races in the population it would be problematic, but a model based on the sample can properly predict vehicle ownership given race. In particular, when the model is multinomial logit, or MNL (as it is here), Manski and Lerman (1977) showed that under certain conditions, the MNL parameter estimates obtained from a stratified sample will be consistent and unbiased relative to the MNL estimates obtained from a simple random sample. Thus, we do not expect that the estimated effect of exposure measures on the number of vehicles owned by the household will be impacted by the biases in our estimation database.

Table 1: Descriptive statistics of quasi-continuous variables.

Quasi-Continuous Variables	Source	Mean	Std. Dev.	1%	99%	Notes
Household adults	Targeted marketing	1.8	0.77	1	4	
Household children	Targeted marketing	0.49	0.92	0	4	
Income (USD)	Targeted marketing	80,196	45,488	10,000	225,000	Given in ranges, bottom-coded at \$10k and top-coded at \$250k. We use the median of the range.
Householder age	Targeted marketing	50	13	26	82	The age in years of the primary adult in the household.
Current Density	Census	299	258	20	1,183	The number of housing units in a ZIP code divided by area measured in square kilometers.
Exposure: Density						
Duration $\alpha = 1$	Calculated using prior adresses from credit reports and density from Census data.	426	813	16	3,242	Each prior residence receives weight proportional to length of residence.
Duration $\alpha = 0$ *(Mean)*		424	716	19	3,100	Each prior residence receives equal weight.
Decay		426	826	16	3,245	Older residences receive less weight than newer ones.
Extreme		704	1,537	24	7,064	The highest density to which the householder has been exposed.
Longest		430	989	8.9	3,393	The density of the ZIP code with the longest time of residence.
Current Alternative Share	Census	0.12	0.054	0.051	0.28	The percentage of work commuters who did not drive alone or carpool.
Exposure: Alt. Share						
Duration $\alpha = 1$	Calculated using prior adresses from credit reports and alternative shares from Census data.	0.11	0.083	0.032	0.46	Each prior residence receives weight proportional to length of residence.
Duration $\alpha = 0$ *(Mean)*		0.11	0.075	0.031	0.41	Each prior residence receives equal weight.
Decay		0.11	0.083	0.031	0.47	Older residences receive less weight than newer ones.
Extreme		0.15	0.13	0.035	0.76	The highest alternative share to which the householder has been exposed.
Longest		0.11	0.096	0.031	0.55	The alternative share in the ZIP code with the longest time of residence.

Table 2: Descriptive statistics of categorical variables.

Discrete Variables	Source	Number	%
Vehicles	Georgia DMV		
1 vehicle		69,899	31
2 vehicle		91,522	40
3+ vehicles		66,409	29
Ethnicity	Targeted marketing		
White		169,758	75
African-American		33,893	15
Asian		5,376	2.4
Hispanic		7,960	3.5
Other		10,843	4.8
Housing Tenure	Targeted marketing		
Owner		214,194	94
Renter		2,645	1.2
Probable Owner		2,585	1.1
Probable Renter		8,406	3.7

2.4 Empirical Model

The number of vehicles y_i that household i owns is assumed to be a function of the need for vehicles N_i, the ability to acquire vehicles A_i, and the individual preference for owning vehicles π_i.

$$y_i = f(N_i, A_i, \pi_i) \tag{6}$$

In application, each of these elements is represented by one or more predictor variables: household size and residential land use may serve as proxies for N_i, whereas A_i mostly consists of income. The attributes of past residences, and other variables such as ethnicity and age, are a means of characterizing π_i.

For cases where y represents finite or discrete outcomes, it is natural to model the *probability* that y takes on a given value, as an ordinal response model or a discrete choice model, such as a multinomial logit model (MNL) (McFadden, 1974). Several prior studies that have compared ordinal response and MNL models for vehicle ownership have found

the MNL models to be superior (Potoglou and Susilo, 2008; Bhat and Pulugurta, 1998); we thus follow this convention and use MNL models. In this case, the f function represents the utility of owning v_i vehicles, and the parameters of that function (as well as, potentially, some of the explanatory variables) can differ with y. More formally, in the MNL, the utility V for household i in choosing alternative j from choice set J is a linear function of $\mathbf{x}_{ij}$, $V_{ij} = \mathbf{x}_{ij}\beta_j + \epsilon_{ij}$, where $\mathbf{x}_{ij}$ comprises the N, A, and π variables described above. If ϵ_{ij} is distributed independently and identically Gumbel (or extreme value type I), the probability of individual i choosing alternative j is given as:

$$P(y = j|\mathbf{x}_{ij}) = \frac{e^{V_{ij}}}{\sum_{k \in J} e^{V_{ik}}} \tag{7}$$

We estimate the MNL model using the `mlogit` package for R (Croissant, 2011). As a measure of model fit, we use the McFadden likelihood ratio index with respect to constants,

$$\rho_C^2 = 1 - \frac{\log(\mathcal{L}_\beta)}{\log(\mathcal{L}_C)} \tag{8}$$

where $\mathcal{L}_\beta$ is the model likelihood, and $\mathcal{L}_C$ is the likelihood of a constants-only (market share) model.

2.5 Results

Our presentation of results proceeds as follows. First, we present results for a base model that includes only current address attributes, representing the most common situation in which only those attributes are available to the analyst. Next, we compare these results to a set of models that each add one of the past exposure metrics developed in Section 2.3.5 to the base model. These specifications allow us to assess if the interpretation of results is robust to different exposure metrics. We use the results from the best fitting model that includes current and past exposure metrics to isolate the autonomous effect of the built environment on vehicle ownership. The last section assesses the relative influence of current and prior exposure on vehicle ownership decisions. A sensitivity analysis of these models is given in Appendix A.

2.5.1 Base Model

Table 3 presents a base model that captures the influences of demographic characteristics, current density and alternative mode share on vehicle ownership. Having insufficient vehicles in a household brings a fairly substantial disutility, as shown by the strongly negative coefficient value (the greater the number of adults who would be without cars — which will be larger for smaller numbers of vehicles — the greater the disutility). Each additional child increases the probability of owning multiple vehicles. An increase[5] in household income is correlated with an increase in the probability of owning more than one vehicle. All else equal, African Americans own fewer cars than Whites, and Whites own fewer vehicles than Hispanics, Asians, and Other ethnic groups. Households that rent are more likely to own fewer vehicles. There is a parabolic effect of householder age: the utility of owning multiple vehicles increases with age to a point (32 years old for two cars, 54 for three or more), and then declines again as the head of household ages. Finally, the density and the alternative mode share associated with the household's current ZIP code show their expected negative coefficients. As density or the number of people commuting without a car increases, the probability of owning two or three or more cars decreases. This finding supports many previous studies (Bhat and Guo, 2007; Cao et al., 2007b; Giuliano and Dargay, 2006) that have shown individuals living in dense areas with transportation alternatives tend to own fewer vehicles, all else equal.

2.5.2 Models that Include Past Exposure Metrics

The models presented in Table 4 control for learned preference and/or self-selection effects by introducing past exposure metrics. The models include the same variables shown in Table 3; however, no coefficients changed by an order of magnitude or by a level of significance[6] and are therefore not shown with the exception of the current density and alternative mode share. The fact that no coefficients meaningfully changed between the base model

[5]The natural logarithmic transformation models a constant effect of a given *percentage* change in income, representing a diminishing marginal effect of each additional dollar of income on utility. The same logic applies to density and alternative share.

[6] We define four significance levels: $p < 0.10, p < 0.05, p < 0.01$, and $p < 0.001$.

Table 3: Base vehicle ownership model.

Generic Variables[1]	β	t-stat
Insufficiency *(# Adults without vehicles)*	-1.088^{***}	-146.0

	2 Vehicles		3+ Vehicles	
Alternative-Specific Variables	β	t-stat	β	t-stat
(Intercept) *ref. 1 vehicle*	-4.236^{***}	-31.1	-10.496^{***}	-64.7
Number of Children	0.127^{***}	19.5	0.071^{***}	9.9
log(Income)	0.371^{***}	39.4	0.452^{***}	42.4
Ethnicity: *ref. White*				
African-American	-0.088^{***}	-5.5	0.050^{**}	2.8
Asian	0.426^{***}	11.5	0.725^{***}	18.5
Hispanic	0.318^{***}	10.4	0.658^{***}	20.3
Other	0.205^{***}	8.1	0.171^{***}	6.0
Housing Tenure: *ref. Known owner*				
Known renter	-0.623^{***}	-13.3	-0.852^{***}	-13.4
Probable renter	-0.568^{***}	-11.5	-1.077^{***}	-12.9
Probable owner	0.080^{**}	2.9	0.281^{***}	9.2
Householder Age	0.012^{***}	5.0	0.186^{***}	54.3
Householder Age^2	$-1.92e{-}04^{***}$	-8.3	$-1.73e{-}03^{***}$	-53.8
log(Current Density)	-0.119^{***}	-14.2	-0.212^{***}	-23.0
log(Current Alternative Share)	-0.119^{***}	-7.2	-0.417^{***}	-22.1
Number of Observations	$227,830$			
Null Log-likelihood	$-250,297$			
Constant Log-likelihood	$-247,926$			
Model Log-likelihood	$-225,528$			
ρ_C^2	0.0903			

** significant at $p < .01$; $^{***}p < .001$

[1] Generic variables take a single coefficient value applicable to all alternatives.

and models that include exposure metrics provides evidence that the exposure metrics are largely uncorrelated with the other model variables, or that our sample size is sufficiently large to overcome the econometric problems of collinear predictors. In no instance did the signs of these coefficients change, showing the directional interpretation of the influence of these variables on vehicle ownership is consistent across all of the exposure formulations.

Overall, the four models (representing different exposure score metrics) have ρ_C^2 fit statistics that are very similar, indicating that the experiences the exposure metrics seek to capture are robust to different specifications. However, the model based on the "Extreme" exposure formulation, which simply considers the prior residence with the highest density or alternative share that the head of the household has experienced, fits the data the best (by a slight margin). Multiple duration exposure formulations were estimated by varying the value α from 0 to 1.3 in 0.1 unit increments. Interestingly, among these duration exposure formulations, the one that fit the data the best is the case where $\alpha = 0.1$, or a very lightly weighted average of all of the previous ZIP codes.

Across all four exposure measure models, the current housing unit density and the current alternative share retain their expected signs and remain highly significant. For example, the results from the extreme exposure formulation show that between two households with identical past exposures and all else equal, the one currently living in a neighborhood with higher densities and/or more non-vehicle transportation utilization is significantly less likely than the other to own multiple vehicles.

The influence of past exposure metrics on vehicle ownership is less clear. Prior exposure to higher densities decreases the probability of owning three or more vehicles; however, the same variable has no discernible difference on the probabilities of owning one versus two vehicles, all else being equal. With one exception, the coefficients associated with the past alternative share are all negative; thus, individuals who have been exposed to higher alternative mode shares are less likely to own multiple vehicles. However, only three out of the eight parameter estimates associated with the past alternative share are significant at the 0.05 level, and thus the relative influence of prior exposure to alternative mode shares is modest.

Table 4: Models incorporating exposure metrics.

	Duration ($\alpha = 0.1$)		Decay		Extreme		Longest	
Variables	β	t-stat	β	t-stat	β	t-stat	β	t-stat
2 Vehicles								
log(Current Density)	-0.115^{***}	-12.7	-0.114^{***}	-12.8	-0.111^{***}	-12.6	-0.119^{***}	-13.5
log(Past Density)	-0.004	-0.5	-0.008	-1.0	-0.011	-1.4	0.002	0.3
log(Current Alternative Share)	-0.111^{***}	-6.5	-0.115^{***}	-6.7	-0.110^{***}	-6.5	-0.115^{***}	-6.8
log(Past Alternative Share)	-0.260^{*}	-2.3	-0.129	-1.3	-0.219^{**}	-3.0	-0.119	-1.4
$3+$ **Vehicles**								
log(Current Density)	-0.138^{***}	-13.9	-0.150^{***}	-15.1	-0.140^{***}	-14.4	-0.178^{***}	-18.4
log(Past Density)	-0.147^{***}	-16.3	-0.119^{***}	-14.6	-0.152^{***}	-18.3	-0.056^{***}	-8.6
log(Current Alternative Share)	-0.404^{***}	-20.9	-0.403^{***}	-20.9	-0.397^{***}	-20.8	-0.402^{***}	-21.0
log(Past Alternative Share)	0.054	0.4	-0.074	-0.6	-0.106	-1.2	-0.314^{**}	-3.3
Model Log-likelihood	$-225,258$		$-225,311$		$-225,150$		$-225,426$	
ρ^2_C	0.0914		0.0912		0.0919		0.0908	

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

2.6 Interpretation of Current and Prior Built Environment Effects

The previous section focused on establishing the statistical significance of the association between prior exposure to the built environment and vehicle ownership. Both current and prior exposure metrics were found to be significant in describing vehicle ownership to some degree. These findings can have multiple interpretations in the theoretical context of either learned preferences or self-selection. In this section, we examine these interpretations in the light of our empirical results.

In a self-selection paradigm, households select to live in neighborhoods that enable their preferred transportation behaviors. In a collective sense, if this is true for the household's current neighborhood, then presumably it should be equally true for its prior neighborhoods. In the previous section we added past information to a base model containing current built environment attributes; a self-selection paradigm might suggest the inverse approach of adding current attributes to a model already controlling for the effects of the past. Table 5 presents three models: a model with only current attributes, a model with only past exposure metrics, and a model with both. These models present two important observations. First, the model using the current coefficients only has a significantly better fit than the model with the past coefficients only; the model with both sets has the best fit. Second, both sets of coefficients change when they are placed in a model together, but the past coefficients change more. Both observations suggest that current neighborhood attributes are better independent predictors than past attributes, but that past experiences may have a moderating effect. At any rate, past attributes are not substitutes for current ones; if households were always able to select their residence to match their preferences, this would be unlikely.

These observations can still be explained under a self-selection theory however, if households are likely to be frequently mismatched (with a land use other than what they prefer), or are more likely to be mismatched at their current address than in their past. Residential mismatch occurs when a household would prefer to live in a built environment other than its current residence (relevant studies include Schwanen and Mokhtarian (2004) and Kamruzzaman et al. (2013)). For example, households that prefer to live in a dense neighborhood

Table 5: Comparison of models that include current attributes and/or past exposure metrics.

	Current Only[1]		Past Only[2]		Both[3]	
Variables	β	t-stat	β	t-stat	β	t-stat
2 Vehicles						
log(Current Density)	-0.119^{***}	-14.2			-0.111^{***}	-12.6
log(Past Density)			-0.053^{***}	-7.3	-0.011	-1.4
log(Current Alternative Share)	-0.119^{***}	-7.2			-0.110^{***}	-6.5
log(Past Alternative Share)			-0.231^{**}	-3.2	-0.219^{**}	-3.0
$3+$ **Vehicles**						
log(Current Density)	-0.212^{***}	-23.0			-0.140^{***}	-14.4
log(Past Density)			-0.221^{***}	-28.2	-0.152^{***}	-18.3
log(Current Alternative Share)	-0.417^{***}	-22.1			-0.397^{***}	-20.8
log(Past Alternative Share)			-0.273^{**}	-3.3	-0.106	-1.2
Model Log-likelihood	$-225,528$		$-225,835$		$-225,150$	
ρ^2_C	0.0903		0.0891		0.0919	

* significant at $p < .05$; $^{**}p < .01$; $^{***}p < .001$

[1] This is precisely the Base model from Table 3.

[2] Using the "extreme" exposure metric.

[3] This is precisely the Extreme model in Table 4.

(and have done so elsewhere in the past) may be unable to find suitable housing in Atlanta if the market under-supplies such homes, and are therefore forced to live in a suburban environment and to own vehicles they might not otherwise out of necessity. The household's current situation therefore dictates their behavior, but there is still a residual preference that the household expresses by owning fewer cars than its otherwise identical neighbors. Conversely, the observations would be explained under a learned preferences mechanism by positing that the household's experience in a more dense neighborhood taught it how to operate without as many vehicles as its neighbors who have not had such an experience.

Irrespective of the operating theory, it is of practical importance to evaluate the strengths of the past effect, since prior addresses may not be readily available to planners who need to estimate a vehicle ownership model. It is straightforward to examine the aggregate error resulting from disregarding past exposure's effect on vehicle ownership. We do this by comparing predicted ownership under two scenarios. The first is an "as is" scenario, applying our extreme model directly to the estimation dataset; our model predicts that our sample households will own a total of 449,145 vehicles.[7] The second scenario eliminates the previous exposure effects by setting the exposure metrics equal to the current address attributes for all households; our model under this scenario predicts that our households will own a total of 454,437 vehicles. This represents an aggregate increase of only 1.2%.

One explanation of this small effect lies directly in the self-selection hypothesis: if most households are able to select into their preferred neighborhoods, there will be too few mismatched households to meaningfully change aggregate vehicle ownership. Indeed, our data seem to bear out this explanation. Of the 227,830 observations in our estimation sample, only 778 show an extreme increase in density, currently residing in a neighborhood with a density far exceeding its historical exposure (15^{th} to 85^{th} percentile). On the other side, there are only 2,594 records showing the opposite behavior, who are now living below the 15^{th} percentile with an observed previous exposure above the 85^{th}. If past and current exposure were two independent random variables, we would expect 5,126 households in both

[7] In reality, the sampled households own 482,331 vehicles; the discrepancy is resolved by considering that some households own four or five vehicles.

groups. There is still the possibility for mismatch, however we observe far fewer households that may be mismatched than the 23.6% estimated by Schwanen and Mokhtarian (2004), though their number was not intended to be representative of a population.

Another explanation for the small aggregate effect of previous exposure is that built environment attributes — either current or historical — have an absolutely small effect on vehicle ownership relative to socioeconomic characteristics. We consider a representative household with two white adults of median age and income, who own their home and have no children. This household is expected to own 2.170 vehicles if the head of household has an average exposure score and the current residence is in a neighborhood with a density and alternative share combination that corresponds to the 15^{th} percentile in each metric. However, if this same household lives in a neighborhood with a density and alternative share combination that respectively correspond to the 85^{th} percentiles, it would be expected to own 2.023 vehicles. This corresponds to a decrease in the number of vehicles owned of about 6.8% (an average elasticity of -0.015). Conversely, if the current neighborhood's density and alternative share are held constant at their means and the head of household's historical exposure moves from the 15^{th} to the 85^{th} percentiles, the household would be expected to own 2.147 versus 2.067 vehicles. This corresponds to a decrease in the number of vehicles of only 3.8% (an average elasticity of -0.0076). Though the current attribute elasticity is two times greater than the previous exposure elasticity, it remains very small in absolute terms.

On one hand, this result is somewhat discouraging, as it suggests that exposure to higher densities and alternate transportation options (either currently or in the past) has a relatively modest influence on vehicle ownership decisions. On the other hand, this result is encouraging from a practical perspective, as it suggests that models that include only current exposure information (in addition to standard socioeconomic information) should be able to accurately predict vehicle ownership decisions. Planners do not likely need to compile extensive histories of prior residences in order to accurately forecast vehicle ownership decisions. However, an extensive database of prior addresses does permit a richer analysis and the ability to isolate the autonomous effects of the built environment

on vehicle ownership decisions, as we have done in this study. This can be useful when designing strategies for reducing vehicle ownership in particular groups of individuals, or evaluating different policy mechanisms for reducing vehicle ownership.

2.7 Limitations and Future Directions

There are several limitations of our study. Due to difficulties in compiling a list of households that do not own vehicles, we needed to exclude zero-vehicle households from our analysis. However, a comparison of vehicle ownership and income calculated from the Census data provides additional insights into which types of households are missing from our database. Based on data collected in the 2011 ACS (U.S. Census Bureau, 2012a), 4.9% of households in the 13-county area do not own a vehicle. Of these zero-vehicle households, 69% are in the bottom income quintile and 85% are in the bottom two quintiles. Thus these households are primarily those who cannot afford to own a vehicle, rather than those who may be expressing a preference for a car-free lifestyle. Studying the preferences of the wealthier 15% of zero-vehicle households (representing 0.72% of all households) would be an important direction for future research.

Another limitation is that this is a retrospective study of households that currently live in the Atlanta metropolitan area. Individuals who once lived in Atlanta, but now reside in areas outside Atlanta are not represented in our analysis database. The ideal analysis of learned preferences and self-selection on vehicle ownership would be based on a national longitudinal panel in which changes in vehicle ownership and densities can be directly observed over time. However, we are unaware of any such database that exists in the U.S.

These shortcomings notwithstanding, the empirical models we have estimated indicate that a preference (either learned or innate) for owning more or fewer vehicles appears to be a real and measurable phenomenon independent from a need established by the existing built environment, but one with a negligible aggregate impact. The effect of prior exposure to density and non-vehicle transportation modes is only about half of the effect of the current

density and non-vehicle commute mode share. This finding does not preclude the possibility that people learn or develop preferences by other methods (such as sociodemographic changes, e.g. in household size, or social influence, or even self-introspection), which in turn may influence transportation behaviors. Studying such preferences directly is likely to be a fruitful research endeavor.

There are several ways in which this study can be extended or modified to investigate particular theories related to the influence of the built environment on an individual's preference development. It may be that some cities are more influential on preferences, or that individuals learn their preferences at different stages in life. For example, an exposure score that weights exceptionally dense regions more heavily, or that gives a bonus to the land use individuals experience in college might prove to be better predictors of vehicle ownership than the metrics used in this study. In terms of self-selection, our analysis suggests that in the Atlanta region, individuals are — for the most part — living in neighborhoods that are similar to those they have lived in previously. This may not be the case for regions in which particular land uses are under-supplied in the market. If the supply for dense neighborhoods with alternative transportation options cannot meet the demand, then many individuals who would like to live in these neighborhoods would be unable to do so. In this situation, the influence of prior experiences (e.g., the "extreme" or "one opportunity" the individual had to live in a dense area) may be a stronger reflection of the individual's preference, and thus more influential on predicting current vehicle ownership.

2.8 Conclusions

Transportation planners have identified urban densification as a policy mechanism to reduce vehicle ownership and/or usage. These professionals cite studies that have shown a correlation between high densities and low vehicle ownership. This correlation is by itself an insufficient basis for policy, however. If households moving into newly dense neighborhoods have previously developed a preference for *high* vehicle ownership, they may simply continue to express those preferences in their new neighborhood. Or it may be that the people who do move in are people who would already have owned fewer vehicles anyway,

because that is the lifestyle they have selected to live. Our study considers precisely these questions, by relating a household's past experiences to its current behavior. We have shown that preferences for vehicle ownership are, at least to some extent, a reflection of prior exposure. Households with prior exposure to higher densities and non-automobile transportation modes are less likely to own multiple vehicles, all else equal. However, the measurable effect of the prior exposure to higher densities and non-vehicle shares, at an individual level or aggregated across the region, is modest.

The central policy implication of our study is that proposed increases in density may have a marginally smaller short-run effect on forecasted vehicle ownership than would be predicted by models that did not consider the new residents' prior experiences. Another implication is that achieving low vehicle ownership rates in a new development is more likely with residents who have previously lived in high-density areas, or who have experience with non-automobile transportation modes.

Our models showed that very large changes in current density and transportation mode share would have a small effect on vehicle ownership when compared with other attributes, such as the number of adults in the household. Thus, planners seeking to limit vehicle ownership should consider other policies in addition to increasing density and providing non-automobile alternatives. Other strategies such as increasing registration fees (Chin and Smith, 1997) or restricting the flow of vehicle traffic (Salon, 2009) into certain areas are likely to have a greater effect on reducing vehicle ownership. The best policy agenda is likely to be a multi-faceted approach, of which densification is only a part.

Acknowledgements We would like to thank Brian Stone for helpful comments on an early version of this paper. This project was partially funded by a U.S. Department of Transportation Eisenhower Fellowship.

2.9 References

Adjemian, M.K., Lin, C.Y.C., Williams, J., 2010. Estimating spatial interdependence in automobile type choice with survey data. *Transportation Research Part A: Policy and Practice* 44, 661–675.

Babbie, E.R., 2009. *The Practice of Social Research.* 12th ed., Wadsworth Publishing Company, Belmont, CA.

Bagley, M.N., Mokhtarian, P.L., 2002. The impact of residential neighborhood type on travel behavior: A structural equations modeling approach. *The Annals of Regional Science* 36, 279–297.

Bhat, C.R., Guo, J., 2007. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological* 41, 506–526.

Bhat, C.R., Pulugurta, V., 1998. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological* 32, 61–75.

Binder, S., Macfarlane, G.S., Garrow, L.a., Bierlaire, M., 2014. Associations among household characteristics, vehicle characteristics and emissions failures: An application of targeted marketing data. *Transportation Research Part A: Policy and Practice* 59, 122–133.

Buckeridge, D.L., Glazier, R., Harvey, B.J., Escobar, M., Amrhein, C., Frank, J., 2002. Effect of motor vehicle emissions on respiratory health in an urban area. *Environmental Health Perspectives* 110, 293–300.

Cao, X., Cao, X., 2013. The impacts of LRT, neighbourhood characteristics, and self-selection on auto ownership: evidence from Minneapolis-St. Paul. *Urban Studies* In press.

Cao, X., Mokhtarian, P.L., Handy, S.L., 2007a. Cross-sectional and quasi-panel explorations of the connection between the built environment and auto ownership. *Environment and Planning A* 39, 830–847.

Cao, X., Mokhtarian, P.L., Handy, S.L., 2007b. Do changes in neighborhood characteristics lead to changes in travel behavior? A structural equations modeling approach. *Transportation* 34, 535–556.

Cao, X., Mokhtarian, P.L., Handy, S.L., 2009. Examining the impacts of residential self-selection on travel behaviour: A focus on empirical findings. *Transport Reviews* 29, 359–395.

Chin, A.T., Smith, P., 1997. Automobile ownership and government policy: The economics of Singapore's vehicle quota scheme. *Transportation Research Part A: Policy and Practice* 31, 129–140.

Croissant, Y., 2011. *mlogit: Multinomial Logit Model*. URL: `http://cran.r-project.org/package=mlogit`.

Dieleman, F.M., Dijst, M., Burghouwt, G., 2002. Urban form and travel behaviour: micro-level household attributes and residential context. *Urban Studies* 39, 507 – 527.

Eluru, N., Bhat, C.R., Pendyala, R.M., Konduri, K.C., 2010. A joint flexible econometric model system of household residential location and vehicle fleet composition/usage choices. *Transportation* 37, 603–626.

Ewing, R., Cervero, R., 2001. Travel and the built environment. *Transportation Research Record* 1780, 87–114.

Fujii, S., Kitamura, R., 2003. What does a one-month free bus ticket do to habitual drivers? An experimental analysis of habit and attitude change. *Transportation* 30, 81–95.

Giuliano, G., Dargay, J., 2006. Car ownership, travel and land use: a comparison of the US and Great Britain. *Transportation Research Part A: Policy and Practice* 40, 106–124.

Groves, R.M., 1989. *Survey Errors and Survey Costs*. John Wiley & Sons, New York.

Handy, S., Cao, X., Mokhtarian, P.L., 2006. Self-selection in the relationship between the built environment and walking: Empirical evidence from northern California. *Journal of the American Planning Association* 72, 55–74.

Kamruzzaman, M., Baker, D., Washington, S., Turrell, G., 2013. Residential dissonance and mode choice. *Journal of Transport Geography* 33, 12–28.

Kenworthy, J.R., Laube, F.B., 1996. Automobile dependence in cities: an international comparison of urban transport and land use patterns with implications for sustainability. *Environmental Impact Assessment Review* 16, 279–308.

Kressner, J.D., Garrow, L.A., 2012. Lifestyle segmentation variables as predictors of home-based trips for Atlanta, Georgia, airport. *Transportation Research Record* 2266, 20–30.

Kressner, J.D., Garrow, L.A., 2013. Using big data for travel demand modeling: A comparison of targeted marketing, Census, and household travel survey data. *Working Paper, Georgia Institute of Technology* .

Leonhardt, D., 2013. In climbing income ladder, location matters. *New York Times*, July 7, 2013. URL: `http://www.nytimes.com/2013/07/22/business/in-climbing-income-ladder-location-matters.html`.

Lois, D., López-Sáez, M., 2009. The relationship between instrumental, symbolic and affective factors as predictors of car use: A structural equation modeling approach. *Transportation Research Part A: Policy and Practice* 43, 790–799.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. *Econometrica* 45, 1977–1988.

Matas, A., Raymond, J.L., Roig, J.L., 2009. Car ownership and access to jobs in Spain. *Transportation Research Part A: Policy and Practice* 43, 607–617.

McFadden, D.L., 1974. Conditional logit analysis of qualitative choice behavior, in: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Mokhtarian, P.L., Cao, X., 2008. Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B: Methodological* 42, 204–228.

Norman, J., MacLean, H.L., Kennedy, C.A., 2006. Comparing high and low residential density: Life-cycle analysis of energy use and greenhouse gas emissions. *Journal of Urban Planning and Development* 132, 10–21.

Pinjari, A.R., Bhat, C.R., Hensher, D.A., 2009. Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B: Methodological* 43, 729–748.

Potoglou, D., Susilo, Y.O., 2008. Comparison of vehicle-ownership models. *Transportation Research Record* 2076, 97–105.

Roorda, M., Carrasco, J., Miller, E.J., 2009. An integrated model of vehicle transactions, activity scheduling and mode choice. *Transportation Research Part B: Methodological* 43, 217–229.

Salon, D., 2009. Neighborhoods, cars, and commuting in New York City: a discrete choice approach. *Transportation Research Part A: Policy and Practice* 43, 180–196.

Sanchez, T.W., Stolz, R., Ma, J.S., 2003. *Moving to equity: addressing inequitable effects of transportation policies on minorities*. Technical Report. Civil Rights Project.

Schrank, D., Eisele, B., Lomax, T., 2012. *Urban Mobility Report*. Technical Report. Texas Transportation Institute. College Station, TX.

Schwanen, T., Mokhtarian, P.L., 2004. The extent and determinants of dissonance between actual and preferred residential neighborhood type. *Environment and Planning B: Planning and Design* 31, 759–784.

Sexton, S.E., Sexton, A.L., 2011. Conspicuous conservation: The Prius halo and willingness to pay for environmental bona fides. UC Center for Energy and Environmental Economics Working Paper Series. URL: `http://ecnr.berkeley.edu/vfs/PPs/Sexton-Ste/web/uceee_conspicuous_cons.pdf`.

Stone, B., 2009. Land use as climate change mitigation. *Environmental Science & Technology* 43, 9052–9056.

U.S. Census Bureau, 2000a. Form D-61B: Census 2000 "long" form questionnaire. URL: `http://www.census.gov/dmd/www/pdf/d-61b.pdf`.

U.S. Census Bureau, 2000b. *H001. Housing Units: Census 2000 Summary File 1 (SF1) 100-percent data.* Technical Report. American FactFinder. URL: `http://factfinder2.census.gov/`.

U.S. Census Bureau, 2000c. *P030. Means of Transportation to Work for Workers 16 Years and Older: 2000 Census Summary File 3 (SF3) - Sample Data. All 5-Digit ZIP Code Tabulation Areas fully-or-partially within United States*. Technical Report. American FactFinder. URL: `http://factfinder2.census.gov/`.

U.S. Census Bureau, 2011. *B08101. Means of Transportation to Work by Age: 2011 ACS 5-year estimates. All ZIP Code Tabulation Areas within the United States*. Technical Report. American FactFinder. URL: `http://factfinder2.census.gov/`.

U.S. Census Bureau, 2012a. *American Community Survey 2006-2010 ACS 5-year PUMS files.* Technical Report. American FactFinder. URL: `http://factfinder2.census.gov/`.

U.S. Census Bureau, 2012b. *B25001. Housing Units: 2007-2011 American Community Survey 5-year estimates*. Technical Report. American FactFinder. URL: `http://factfinder2.census.gov/`.

Van Acker, V., Witlox, F., 2010. Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. *Journal of Transport Geography* 18, 65–74.

Weinberger, R., Goetzke, F., 2010. Unpacking preference: How previous experience affects auto ownership in the United States. *Urban Studies* 47, 2111–2128.

CHAPTER III

DO ATLANTA RESIDENTS VALUE MARTA? SELECTING AN AUTOREGRESSIVE MODEL TO RECOVER WILLINGNESS-TO-PAY

Gregory S. Macfarlane, Laurie A. Garrow, Juan Moreno-Cruz

Submitted to *Transportation Research Part A*, 2013

Chapter Abstract

Understanding homeowners' marginal willingness-to-pay (MWTP) for proximity to public transportation infrastructure is important for planning and policy. Naïve estimates of MWTP, however, may be biased as a result of spatial dependence, spatial correlation, and spatially endogenous variables. In this paper we discuss a class of spatial autoregressive models that control for these spatial effects, and apply them to sample data collected for the Atlanta, Georgia housing market. We provide evidence that a general-to-specific model selection methodology that relies on the generality of the spatial Durbin model (SDM) should be preferred to the classical specific-to-general methodology that begins with an assumption of no spatial effects. We show that applying the SDM widens the confidence interval of the estimate of MWTP for transit proximity in Atlanta, relative to ordinary linear regression. This finding has unpredictable consequences for land value capture forecasts and transportation policy decisions.

3.1 Introduction

Home equity accounts for the largest share of household wealth in the United States, representing 78% of the net worth for the median household in 2010 (U.S. Census Bureau, 2010). Correspondingly, property taxes represent the largest independent revenue stream for local governments (Tax Policy Center, 2013). Municipal authorities have at least two interests in

policy mechanisms that are correlated with raising home and property values: the welfare of their citizens and their municipalities' fiscal health.

One strategy that authorities may employ to raise property values is to construct transportation infrastructure. In theory, the rent price at locations with good accessibility to jobs, markets, schools or other activity centers will be higher than at locations with poor accessibility (Alonso, 1960). Advocates of public mass transit in particular point to a strategy of land value capture resulting from transit development (Smith and Gihring, 2006): if a region[1] expends resources to improve the public transit system in a neighborhood, rents in that neighborhood should rise. This public expenditure will increase the private wealth of landholders in the improved area and consequently property tax revenues in the region. Whether government can recoup the cost of its infrastructure expenditure over a reasonable period of time, however, is an empirical question that remains unanswered, because the willingness of households to pay for a marginal improvement in their transportation accessibility is not entirely understood, and is perhaps heavily dependent on local circumstances.

The marginal willingness to pay (MWTP) for a characteristic of a good is a direct function of the utility derived from that characteristic (Rosen, 1974); therefore, a regression model with the price of the good as the dependent variable and the good's characteristics as predictor variables — called a *hedonic* model — reveals the MWTP for each characteristic if the assumptions of regression are met. The residual error terms, for instance, must be independently and identically distributed else the researcher cannot test that the MWTP is not zero. The challenge for researchers who develop hedonic home price models is that characteristics of urban housing markets interfere with regression assumptions in four important ways:

1. **Spatial dependence of prices:** The housing market is comparative; the price of a home is relative to the prices of homes nearby. This creates a missing variable bias in the linear regression model.

[1] In the US, metropolitan planning organizations direct local, state, or federal funds to major transportation investments. Local governments generally collect property taxes. In a land value capture strategy, these disparate levels of government work in concert.

2. **Spatially correlated error terms:** Homes near each other have similar characteristics. This will violate the linear regression assumption that error terms are distributed independently, thus invalidating significance tests.

3. **Spatially endogenous or omitted variables:** Neighborhood attributes that are unobservable, such as neighborhood prestige, raise or lower the prices of homes. This can cause both a missing variable bias and cause correlated errors.

4. **Spatial heterogeneity (non-stationarity):** The housing market in one neighborhood may value some attributes more highly than the market in another. These differing preference are reflected in model parameters that vary in space.

Whereas the first three problems are types of *autocorrelation* and have a similar solution in autoregressive models, the solution to the fourth is to fit locally-weighted regressions as described by Brunsdon et al. (1999). This paper focuses on the first three problems and limits the discussion to autoregressive models. Spatial dependence and correlation are fundamentally different, although spatial endogeneity can be seen as a combination of both. Spatial dependence is a substantive problem, resulting in biased estimates of model parameters. Spatial correlation is a nuisance problem, affecting not the parameter estimates themselves but estimates of their standard errors. Identifying which problems may exist in a particular dataset or hedonic model is an important practical question for transportation researchers, on whose models transportation investment and policy decisions rely.

In this paper, we compare two modeling frameworks that have been used to identify spatial processes in housing markets with a particular emphasis on recovering the MWTP for public transit proximity. The first is the classical framework that relies on Lagrange multiplier tests for spatial dependence and correlation in the linear model residuals (Anselin, 1988b). The second is a general framework that tests for restrictions in a general nesting model (the spatial Durbin model, or SDM). The two frameworks may identify different preferred models in certain circumstances. The use of the classical selection framework in recent studies examining spatial autoregression in an accessibility context (Osland, 2010; Löchl and Axhausen, 2010; Ibeas et al., 2012) may have led the authors to select an inferior

model, in which spatial dependence and spatial endogeneity were not incorporated. We show that applying spatial econometrics to the MWTP problem widens the confidence interval around the expected MWTP relative to a simple linear model. This finding may have implications for risk estimations in land value capture strategies.

The remainder of this section provides the context for spatial econometric models in the larger hedonic evaluation literature. Section 3.2 reviews spatial econometric models and the two frameworks that have been used to select a preferred model. Sections 3.3 and 3.4 apply these two frameworks to the Atlanta region and the MARTA heavy-rail system and discuss results. Finally, Sections 3.5 and 3.6 compare our proposed framework to recent studies that have used one or both of the modeling frameworks and offer perspectives on future research objectives.

3.1.1 Home Prices and Transit Accessibility

Some of the earliest hedonic home price models showed a strong relationship between transportation network accessibility and home prices. Brigham (1965) showed that the accessibility potential of a parcel, defined by its access to highway networks, was a better predictor of home values in Los Angeles than its distance to the central business district. Dubin and Sung (1987) observed a similar result in Baltimore. Other early studies used highway accessibility as a control variable in a more holistic model of housing markets (Massell and Stewart, 1971; Ridker and Henning, 1967).

Researchers in the last thirty years have been particularly interested in the hedonic value of transit proximity, as many cities have opened or expanded public rail transit networks. Simple linear hedonic models abound, and generally show a positive MWTP for transit accessibility (Grass, 1992; Lewis-Workman and Brod, 1997). Other researchers have segmented their data or introduced variables to examine particular theories. Chen et al. (1998), for instance, showed that transit stations in Portland have both a positive proximity benefit on prices and also a negative nuisance effect stemming from the increase mechanical noise and foot traffic around stations. Nelson (1992) observed two distinct MWTP estimates in Atlanta, with lower-income neighborhoods valuing transit proximity more than

higher-income neighborhoods.

Urban housing markets are complicated systems, and researchers have applied numerous econometric techniques to isolate MWTP for transit proximity from other confounding variables. Bowes and Ihlanfeldt (2001) estimated sub-models of crime rate and retail activity in areas around Atlanta transit stations, and then used the predictions from these models as instruments in a hedonic model; this process may remove econometric endogeneity from variables that influence home prices but that are only indirectly related to the transit station. A number of studies have used time series or panel data methods to eliminate the effects of unobserved variables — with the assumption that these unobserved or endogenous variables do not change over time — and establish the direct effect of improved transportation accessibility on home prices (Chernobai et al., 2009; McMillen and McDonald, 2004; Mikelbank, 2004; Iacono and Levinson, 2011). These methods may be less applicable to cities with mature transportation networks, where the transportation network is fixed and other variables in the housing market are changing instead.

The complexity of urban housing markets calls for econometric models that can provide unbiased estimates of MWTP for home or location characteristics, while still allowing for parsimonious model specification (Dubin et al., 1999). Numerous elements of a home or its neighborhood might influence its price, and to expect any researcher to capture all of these elements in a data vector is unrealistic. Rather than expand the variables included in an econometric model, the field of spatial econometrics leverages Tobler's first law of geography, that "nearer things are more related than distant things" (Tobler, 1970). By parsimoniously representing these relationships, spatial econometric models can produce unbiased measures of MWTP across a wide range of policy variables.

3.2 Methodology

The linear regression model, which is the traditional starting point for hedonic models, is expressed as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{9}$$

where $\boldsymbol{y}$ is a vector of length n (the number of observations) and X is an $n \times p$ matrix of p attributes (including a constant intercept term). The average marginal effect of the attributes $\boldsymbol{x}_k \in X$ on $\boldsymbol{y}$ is given by the vector of slope parameters $\boldsymbol{\beta}$, which has an ordinary least-squares (OLS) estimator $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y} \tag{10}$$

If the mean of the residuals $\boldsymbol{\epsilon}$ is 0, then the expected value of this estimator is unbiased. The variance of this estimator is $\sigma^2(X'X)^{-1}$, under the Gauss-Markov assumption that $\boldsymbol{\epsilon}$ is distributed independently and identically with a normal distribution of mean zero and variance σ^2; mathematically, $\text{Var}\,(\boldsymbol{\epsilon}|X) = \sigma^2 I$. This assumption underlies all of the standard hypothesis tests on the significance of the elements of $\boldsymbol{\beta}$.

Two basic situations (among many) are known to interfere with OLS estimation. Variables that are excluded from X but which are nonetheless important to $\boldsymbol{y}$ will lead to biased estimates of $\boldsymbol{\beta}$. Error terms that are not independently or identically distributed will produce a biased estimate of σ, thereby invalidating hypothesis tests. Spatial dependence is effectively an omitted variable problem; a home's value depends at least partially on the values of nearby homes, and these values should therefore be incorporated into X. Spatial correlation, on the other hand, is econometrically similar to heteroskedasticity in that it creates unreliable estimates of model standard error.

3.2.1 Spatial Autoregressive Models

Spatial autocorrelation in a variable $\boldsymbol{x}$ may be represented by the relationship $\boldsymbol{x} = \rho W\boldsymbol{x}$, where ρ is a correlation coefficient and W is an $n \times n$ spatial weights matrix that maps each x_i onto its "neighbors" $x_j, j \in 1 \ldots n$. Elements w_{ij} of W are zero if i and j are not neighbors and positive if i and j are neighbors (more detail on spatial weights matrices is given in Dubin (1998). The correlation coefficient ρ is a measure of the strength of the spatial autocorrelation within $\boldsymbol{x}$; values of ρ close to zero indicate that there is little spatial autocorrelation in $\boldsymbol{x}$, and values close to one indicate that there is strong spatial autocorrelation in $\boldsymbol{x}$.

If the dependent variable $\boldsymbol{y}$ is autocorrelated, then the sample exhibits spatial dependence. A model that attempts to replicate this data generation process is the Cliff and Ord (1970) *spatial simultaneous autoregressive lag model* (SAR):

$$\begin{aligned} \boldsymbol{y} &= \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{y} &= (I - \rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \end{aligned} \tag{11}$$

This model, which contains a *spatial lag* of the dependent variable, will provide estimates of $\boldsymbol{\beta}$ that are robust to spatial dependence, provided that the researcher's assignation of W is sufficiently close to the true, unobserved spatial structure. The corresponding model that addresses spatially autocorrelated errors is the *spatial error model* (SEM),

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\beta} + \boldsymbol{u}, \boldsymbol{u} = \lambda W \boldsymbol{u} + \boldsymbol{\epsilon} \\ \boldsymbol{y} &= X\boldsymbol{\beta} + (I - \lambda W)^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{12}$$

where λ is the correlation coefficient for the errors. Again, if W is sufficiently close to the true spatial relationship, the SEM will produce estimates of model standard error that are robust to residual autocorrelation.

A third model, the *spatial Durbin model* (SDM), was originally derived by Anselin (1980) as a consolidated form of the SEM

$$\boldsymbol{y} = \lambda W \boldsymbol{y} + X\boldsymbol{\beta} + WX(-\lambda\boldsymbol{\beta}) + \boldsymbol{\epsilon} \tag{13}$$

This model has spatial lags of both the dependent and the independent variables. The SDM may also be estimated in an unrestricted form by allowing the lagged independent variable parameter vector $-\lambda\boldsymbol{\beta}$ to take its own maximum likelihood value $\boldsymbol{\gamma}$ (Burridge, 1981).

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{14}$$

The unrestricted SDM is a linear combination of the SAR and SEM, and therefore is robust to both spatial dependence and spatial correlation.

The SDM has two further econometric properties that make it particularly appropriate for hedonic analysis. First, because the model contains a set of lagged predictors $\boldsymbol{\gamma}$, it

explicitly models the externality that attributes of observation j impose on the outcome for observation i (Anselin, 2003). A consequence of this property is that the marginal effect of $\boldsymbol{x}_k$ on $\boldsymbol{y}$ is *not* β_k as in a linear model. Instead, there exist separate direct, indirect, and total effects that the analyst must consider. The theoretical average effects $M(k)$ of a variable $\boldsymbol{x}_k$ on $\boldsymbol{y}$ in a SDM are

$$\begin{aligned} M(k)_{\text{direct}} &= n^{-1} tr((I - \rho W)^{-1}(I\beta_k + W\gamma_k)) \\ M(k)_{\text{total}} &= n^{-1} \iota'(I - \rho W)^{-1}(I\beta_k + W\gamma_k)\iota \\ M(k)_{\text{indirect}} &= M(k)_{\text{total}} - M(k)_{\text{direct}} \end{aligned} \tag{15}$$

where ι is a vector of ones of length n. It is important to note that all three types of effects are linear functions of β_k, γ_k, W, and ρ. Details on efficiently calculating these effects are given by LeSage and Pace (2009, p. 114). Further, a Monte Carlo analysis of the effects with draws of β_k, γ_k, and ρ based on the analytical model parameter variance-covariance matrix can produce empirical standard errors of the effects.

This explicit modeling of direct and indirect effects has made the SDM popular in modeling systems where externalities are important, such as the home price penalty of being downwind from swine farms (Kim and Goldsmith, 2008). The SDM is also especially appropriate for systems where the observations interact with each other, such as trade between regions (LeSage and Fischer, 2008). Whether the analyst pays more attention to direct, indirect, or total effects will depend on her specific problem; recommended policy interventions are dependent on which type of effect is considered. In the case of MWTP for transit accessibility, we are mostly concerned with the total effect.

The second econometric property of the SDM is that it may be seen to arise from an autoregressive fixed effects process, and can therefore control for spatially correlated endogenous or omitted variables, given some assumptions. Neighborhood prestige, for example, is not an attribute that can be measured properly; school quality or crime rates may play a role, but these variables are endogenous in that quality neighborhoods create quality schools, and vice-versa. Assume that each observation i inherits some unobserved fixed effects based on its spatial location a_i. We wish to include the entire vector of fixed

effects $\boldsymbol{a}$ (a vector of length n) in a model to control for the endogenous omitted variables and to remove all correlation between X and ε, but such a model would be inestimable as it would contain $n + p$ variables and only n observations. If we assume, however, that the fixed effects $\boldsymbol{a}$ follow a spatial autoregressive process and are correlated with the X terms, we can construct the data generating process

$$\boldsymbol{a} = \rho W \boldsymbol{a} + X\boldsymbol{\gamma}' + \boldsymbol{\epsilon} \tag{16}$$

with ρ a correlation coefficient, $\boldsymbol{\gamma}'$ a vector of estimable parameters of length p, and $\boldsymbol{\epsilon}$ assumed to be distributed IID normal. Solving Equation 16 for $\boldsymbol{a}$ yields an expression for the fixed effects

$$\boldsymbol{a} = (I - \rho W)^{-1}(X\boldsymbol{\gamma}' + \boldsymbol{\epsilon}) \tag{17}$$

Replacing the error of the linear model with the spatial autoregressive fixed effect given in Equation 17 and rearranging terms,

$$\begin{aligned}
\boldsymbol{y} =& X\boldsymbol{\beta} + (I - \rho W)^{-1}(X\boldsymbol{\gamma}' + \boldsymbol{\epsilon}) \\
(I - \rho W)\boldsymbol{y} =& (I - \rho W)X\boldsymbol{\beta} + X\boldsymbol{\gamma}' + \boldsymbol{\epsilon} \\
\boldsymbol{y} =& \rho W\boldsymbol{y} + X(\boldsymbol{\beta} + \boldsymbol{\gamma}') + WX(-\rho\boldsymbol{\beta}) + \boldsymbol{\epsilon} \\
\boldsymbol{y} =& \rho W\boldsymbol{y} + X\boldsymbol{\beta}^* + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

gives the SDM presented in Equation 13 (LeSage and Pace, 2009, p. 29). This analysis implies that the effects of neighborhood variables that are omitted or unobservable, such as school quality, crime rates, or neighborhood prestige are controlled for with the stipulation that these variables themselves follow a spatial autoregressive process. If, on the other hand, the missing variables are spatially uncorrelated or exogenous, then spatial models may be unnecessary.

3.2.2 Model Selection

Table 6 presents a summary of the consequences for using a mis-specified spatial model. A failure to account for spatial dependence, by using an OLS or SEM when an SAR or SDM is the true model, results in biased estimates of the model parameters. Failure to

Table 6: Consequences of misspecified hedonic model.

	Estimated Model			
True DGP	OLS	SAR	SEM	SDM
OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$	-	inefficient	inefficient	inefficient
SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\hat{\boldsymbol{\beta}}$ biased	-	$\hat{\boldsymbol{\beta}}$ biased	inefficient
SEM:$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$	$\hat{\sigma}^2$ invalid	$\hat{\sigma}^2$ invalid	-	inefficient
SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{u}$	$\hat{\boldsymbol{\beta}}$ biased	$\hat{\sigma}^2$ invalid	$\hat{\boldsymbol{\beta}}$ biased	-

account for correlated model errors, by using an OLS or SAR when an SEM or SDM is the true model, will result in biased estimates of standard errors and the invalidation of parameter significance tests. Using any spatial model when it is not required reduces the efficiency of the estimates, as the analyst will sacrifice degrees of freedom to estimate unneeded parameters. Testing whether ρ or λ are zero in the estimates of Equation 11 or Equation 12 is direct but inadequate, as spatial correlation may appear to be spatial dependence and vice versa; a more robust selection framework is required. There are two primary frameworks that analysts may use to select the appropriate spatial autoregressive model.

3.2.2.1 Classical Framework

Least-squares estimates of spatial autoregressive models are inconsistent and the models must therefore be estimated using maximum likelihood techniques. The log-likelihood function for the SAR model (for example) is (Anselin, 1988b):

$$\ln(\mathcal{L}_{SAR}) = -(n/2)\ln(\pi\sigma^2) + \ln|I - \rho W| - \frac{e'e}{2\sigma^2}, \tag{18}$$
$$e = \boldsymbol{y} - \rho W\boldsymbol{y} - X\boldsymbol{\beta}$$

Calculating the determinant of an arbitrary $n \times n$ matrix is a computationally expensive process of $O(n!)$, and the log-determinant term $\ln|1 - \rho W|$ is present in the likelihood functions for the SAR, the SEM, and the SDM models. There are features of W that reduce the computational order (LeSage and Pace, 2009, Chap. 4), but even on a modern computer this is a time-consuming process. When these models were first developed, there

was an enormous incentive to find tests for autocorrelation that avoided computing the likelihood function in Equation 18. The classical framework was born from the need to avoid intensive computer calculations.

Anselin (1988a) developed Lagrange multiplier tests for spatial dependence (LM_ρ) and spatial correlation (LM_λ) that are estimated using the OLS residuals, $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$:

$$LM_\rho = \frac{(\hat{\boldsymbol{\epsilon}}' W \boldsymbol{y}/\hat{\sigma}^2)^2}{nJ} \tag{19}$$

$$LM_\lambda = \frac{(\hat{\boldsymbol{\epsilon}}' W \hat{\boldsymbol{\epsilon}}/\hat{\sigma}^2)^2}{T} \tag{20}$$

with

$$J = \frac{1}{n\hat{\sigma}^2}[(WX\hat{\boldsymbol{\beta}})'(I - X(X'X)^{-1}X')(WX\hat{\boldsymbol{\beta}}) + T\hat{\sigma}^2] \tag{21}$$

and T the trace of the matrix $W'W + W^2$. Both of these statistics are asymptotically distributed χ^2_1. Rejecting the null hypothesis that $LM_\rho = 0$ implies that the proper model is the SAR. Similarly, rejecting that $LM_\lambda = 0$ implies that the proper model is the SEM.

The LM tests suffer from the same confusion as the direct parametric tests on λ and ρ; that is, spatial dependence can appear as spatial correlation in the model residuals and *vice versa*. It is therefore not uncommon that *both* LM tests will reject the null hypothesis. For this reason, Anselin et al. (1996) proposed *robust* LM tests:

$$RLM_\rho = \frac{(\hat{\boldsymbol{\epsilon}}' W \hat{\boldsymbol{\epsilon}} - T(nJ)^{-1}\hat{\boldsymbol{\epsilon}}' W \boldsymbol{y}/\hat{\sigma}^2)^2}{nJ - T} \tag{22}$$

$$RLM_\lambda = \frac{(\hat{\boldsymbol{\epsilon}}' W \boldsymbol{y} - \hat{\boldsymbol{\epsilon}}' W \hat{\boldsymbol{\epsilon}}/\hat{\sigma}^2)^2}{T[1 - T(nJ)]^{-1}} \tag{23}$$

It is even possible that both of the RLM statistics will reject their null hypotheses; in this case, Florax et al. (2003) recommend selecting the model (either SAR or SEM) with the larger test statistic. This framework does not lead to the unrestricted SDM model, though in the presentation by Osland (2010), the SDM should be used if the robust LM tests are "inconclusive." This selection framework is described in Algorithm 1.

3.2.2.2 General Framework

An alternative strategy that we term the "general" framework, was proposed by Florax et al. (2003), and relies on the fact that the SDM is a linear combination of the SAR and

Algorithm 1 Classical Selection Framework

procedure SpEffects($\boldsymbol{y}, X, W$)
  Obtain $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y}$ ▷ OLS residuals
  Calculate $LM_\rho : \rho \overset{?}{=} 0$ AND $LM_\lambda : \lambda \overset{?}{=} 0$ ▷ Lagrange multiplier tests
  if $\rho = 0$ AND $\lambda = 0$ **then**
    OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ Efficient, risk bias in $\boldsymbol{\beta}, \sigma$
  else if $\rho \neq 0$ AND $\lambda = 0$ **then**
    SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
  else if $\rho = 0$ AND $\lambda \neq 0$ **then**
    SEM: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
  else
    Calculate $RLM_\rho : \rho \overset{?}{=} 0$ AND $RLM_\lambda : \lambda \overset{?}{=} 0$ ▷ Robust LM tests
    if $\rho \neq 0$ AND $\lambda = 0$ **then**
      SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
    else if $\rho = 0$ AND $\lambda \neq 0$ **then**
      SEM: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
    else if $\rho \neq 0$ AND $\lambda \neq 0$ **then**
      if $RLM_\rho > RLM_\lambda$ **then**
        SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
      else if $RLM_\lambda > RLM_\rho$ **then**
        SEM: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
      else
        SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}, \sigma$ unbiased, risk inefficiency
      end if
    end if
  end if
end procedure

SEM. Previously, Hendry (1979) proposed that whenever a general nesting model (such as the SDM) exists, it is appropriate to begin the specification search there. LeSage and Pace (2009) argue that from a Bayesian model uncertainty perspective, this alternative strategy is the only appropriate approach (page 31). Consider the SDM ($\boldsymbol{y}_c$) as a weighted linear combination of the SAR ($\boldsymbol{y}_a$) and SEM ($\boldsymbol{y}_b$) models,

$$\begin{aligned}
\boldsymbol{y}_c &= \pi_a \boldsymbol{y}_a + \pi_b \boldsymbol{y}_b \\
\boldsymbol{y}_c &= \pi_a((I-\rho W)^{-1}(X\boldsymbol{\beta}+\boldsymbol{\epsilon})) + \pi_b(X\beta + (I-\rho W)^{-1}\boldsymbol{\epsilon}) \\
(I-\rho W)\boldsymbol{y}_c &= X(\pi_a\boldsymbol{\beta}) + (I-\rho W)(X\pi_b\boldsymbol{\beta}) + (\pi_a+\pi_b)\boldsymbol{\epsilon} \qquad (24) \\
\boldsymbol{y}_c &= \rho W \boldsymbol{y_c} + (\pi_a+\pi_b)X\boldsymbol{\beta} + WX(-\rho\pi_b\boldsymbol{\beta}) + \boldsymbol{\epsilon} \qquad (25)
\end{aligned}$$

with π_a the probability of an SAR and π_b the probability of an SEM, $\pi_a + \pi_b = 1$. If the true specification is an SAR, then no data evidence will show that $\pi_b > 0$, and Equation 24 will reduce to the SAR. Conversely, if the true specification is an SEM, then $\pi_a = 0$, and the SEM will remain. For any situation in which there exists uncertainty, $0 < \pi_a, \pi_b < 1$, the full SDM in Equation 25 should be used.

The analyst implements this framework by estimating the SDM and SEM models. If the true model is the SEM, then $\boldsymbol{\gamma} = -\rho\boldsymbol{\beta}$ (from Equation 14), and the two models will have the same model likelihood. A likelihood ratio (LR) test can be used,

$$-2(\ln(\mathcal{L}_{SEM}) - \ln(\mathcal{L}_{SDM})) \sim \chi^2_1 \qquad (26)$$

If, on the other hand, the true model is an SAR, then the lagged independent parameters $\boldsymbol{\gamma} = 0$ and the reduction is trivial. Finally, if $\rho = 0$, then an OLS model is sufficient. Details of this selection framework are given in Algorithm 2.

3.2.2.3 Comparisons of the Two Frameworks

There have been a number of studies comparing the two modeling frameworks. In essentially all of these studies, the researcher creates a known data generating process, and executes a Monte Carlo simulation to recover this DGP. In the previously cited work by Florax et al. (2003), the researchers find that the classical approach identifies the correct data

generating process (which is an SAR, SEM, or OLS model) more frequently than the LR-based general approach. For high values of spatial correlation however (ρ, λ approaching 1), the two methods are shown to have very similar outcomes. It was not possible in this study for the true model to be an SDM.

By contrast, Larch and Walde (2008) showed that when the SDM was an available option the general framework performed better. Further, they recommend that Wald tests for common factors be used instead of LR tests, because their power was higher in the range of autocorrelation tested. The true spatial parameters used by Larch and Walde were very small, however, with $\rho, \lambda \leq 0.2$. This corroborates the results of Mur and Angulo (2006), who compared the power of the LR test statistic for common factors against LM and Wald tests of the same hypothesis. They showed that, with large sample sizes and high R^2 statistics in the OLS model, that the three tests performed equivalently. This implies that the choice of statistic depends on the strength of spatial correlation, but that at higher levels the difference is negligible.

Algorithm 2 General Selection Framework

procedure SpEffects($\boldsymbol{y}, X, W$)
  Estimate SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}, \sigma$ unbiased, risk inefficiency
  if $\boldsymbol{\gamma} = -\rho\boldsymbol{\beta}$ **then**
    SEM:$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
    if $\lambda = 0$ **then** ▷ No correlation
      OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ Efficient, risk bias in $\boldsymbol{\beta}, \sigma$
    end if
  else if $\boldsymbol{\gamma} = \mathbf{0}$ **then**
    SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
    if $\rho = 0$ **then** ▷ No Dependence
      OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ Efficient, risk bias in $\boldsymbol{\beta}, \sigma$
    end if
  else
    SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
  end if
end procedure

3.3 Empirical Application

3.3.1 Data

This study used targeted marketing (TM) records, which are maintained by credit reporting agencies and other private firms in an effort to assess the creditworthiness of adults in the United States. TM records are compiled from bank records, credit card statements, and other sources both public and private, and contain detailed information on household demographics, financial obligations, and consumption patterns. TM records represent an emerging and potentially important source of data for transportation research (Kressner and Garrow, 2012). A sample of TM records representing the 13-county Atlanta non-attainment area was joined to transportation network shapefiles available from the Atlanta Regional Commission (2012). We restricted the analysis to a cross-section of owner-occupied homes purchased in 2009 and 2010 and located within five miles of a Metropolitan Atlanta Rapid Transit Authority (MARTA) heavy rail transit station. Using a cross-section of sales from a small time frame eliminates the need to consider temporal dependence; spatial autoregressive models can accommodate temporal effects, but LeSage and Pace (2009, chap. 2) also show that a cross-sectional spatial autoregressive model represents the equilibrium point of a temporally dependent process. The MARTA system operates in only two of metropolitan Atlanta's 13 counties; limiting the scope of the analysis to homes within five miles of a MARTA station avoids confusing proximity to MARTA with proximity to central Atlanta, two measurements which will be highly collinear for homes near the periphery of the region.

The 4,812 observations, mapped in Figure 2, match our expectation of settlement patterns in the city. In the northern part of the study area a number of observations are missing in a shape corresponding to the boundaries of the City of Sandy Springs. The similarly large empty space in the south corresponds to a military base and Atlanta Hartsfield-Jackson International Airport. Descriptive statistics for the sample are given in Table 7. Analysis available from the authors shows that this sample does not deviate materially from data for the Atlanta region collected through the US Census Bureau's American Community Survey.

3.3.2 Model

We predict the price of a home as a function of its proximity to a rail station, conditioned on attributes of the home. The basic regression model is semi-log,

$$\log(\mathit{Value})_i = f(\boldsymbol{H}_i, \boldsymbol{O}_i, \boldsymbol{A}_i) \tag{27}$$

where $\boldsymbol{H}_i$ is a vector of house attributes including the square footage and property acreage of a home, its structural type (single dwelling, multiple dwellings, or mobile), and its age. $\boldsymbol{O}_i$ is a vector of attributes describing the homeowner, namely the household annual income and the ethnicity of the householder.

$\boldsymbol{A}_i$ describes the transportation accessibility of the home, which for our study is the natural logarithm of the Euclidean distance (in miles) between the home and its nearest MARTA station. The existing literature on hedonic models of transportation accessibility uses a wide array of specifications, including linear distance (Grass, 1992), binary proximity (Bowes and Ihlanfeldt, 2001), linear distance conditioned on binary proximity (Hess and Almeida, 2007), and spline regression (Chernobai et al., 2009). Proximity is an inherently continuous phenomenon: a home 0.51 miles from a rail station is only 100 feet from a home that is 0.49 miles from the station, a surely negligible distance for virtually all homebuyers. Further, some homebuyers may feel they have access at 0.75 miles, and others feel they do not past 0.25 miles. This fact is highlighted by Debrezion et al. (2007), who find continuous functions generally are better predictors of residential property values than proximity dummies (though they find the reverse is true for commercial properties). Using the natural logarithm applies a diminishing marginal cost to the distance; that is, 100 feet proportionally adds more to the cost of a 400-foot journey than to a journey of 2,000 feet. This same logic of diminishing marginal cost or returns compels us to similarly log-transform the home value, household income, square footage, and property acreage for the observations. Highway accessibility also should affect home prices; we control for this by including the natural logarithm of the distance in miles to the nearest freeway entrance point.

Table 7: Descriptive statistics of model variables

Continuous Variables	Mean	Median	Std. Dev.	Min	Max
Market value of home (kUSD)	311	225	274	50	2,000
Home built area (square feet)	2,125	1,875	1,115	750	7,000
Property area (acres)	0.452	0.375	1.22	0.25	40
Age of home (years)	38.2	38	24.8	1	120
Household Income (kUSD)	84.5	62.5	56.8	10	250
Distance to MARTA rail station (miles)	2.24	2.03	1.39	0.0403	5.21
Distance to freeway entrance (miles)	1.4	1.26	0.835	0.0767	4.61

Discrete Variables	Number	%
Property type		
Single dwelling unit	4,318	87.8
Multiple dwelling units (condominium)	599	12.2
Ethnicity		
White	2,644	53.8
African-American	1,567	31.9
Asian	175	3.56
Hispanic	144	2.93
Other	387	7.87

3.3.3 Spatial Weights

For the autoregressive models in this study, houses were neighbors if they were located within 1.8 miles of each other. The link was weighted by the inverse distance between neighbors to give more consideration to nearer observations. Our neighbors matrix W is the row-standardized inverse Euclidean distance between observations

$$W_{ij} = \begin{cases} \frac{1/d_{ij}}{\sum_{k=1}^{n}(1/d_{ik})} & \text{for } d_{ij} \leq 1.8 \text{ miles} \\ 0 & \text{for } d_{ij} > 1.8 \text{ miles} \end{cases} \tag{28}$$

Row-standardization is a standard technique that aids in interpretation (LeSage and Pace, 2009). This weighting scheme was selected because it provided the maximum model likelihood in a comparison of different schema and radii. A full description of this selection procedure is given in Appendix B.

3.4 Model results

Maximum likelihood estimates of the OLS, SAR, SEM, and SDM models were calculated using the "spdep" package for R (Bivand, 2006; R Development Core Team, 2013); the parameter estimates and statistics are given in Table 8.

Classical Selection The results of the LM tests are presented in Table 9. As shown, we reject the null hypothesis that there is no spatial dependence or spatial correlation in the model. As the RLM_λ statistic is an order of magnitude larger than the RLM_ρ statistic, the SEM is conclusively the preferred model.

General Selection We reject the null hypothesis that the SDM and the SEM fit the data equally well, as the SDM produces a significantly higher model likelihood (p-value of 0). And as not all of the lagged parameters $\boldsymbol{\gamma} = 0$, the SDM remains the most appropriate model. This disagreement between the two selection methodologies is concerning, and highlights the importance of the selection framework. In this case in particular, the choice of model has significant implications for the model outcomes.

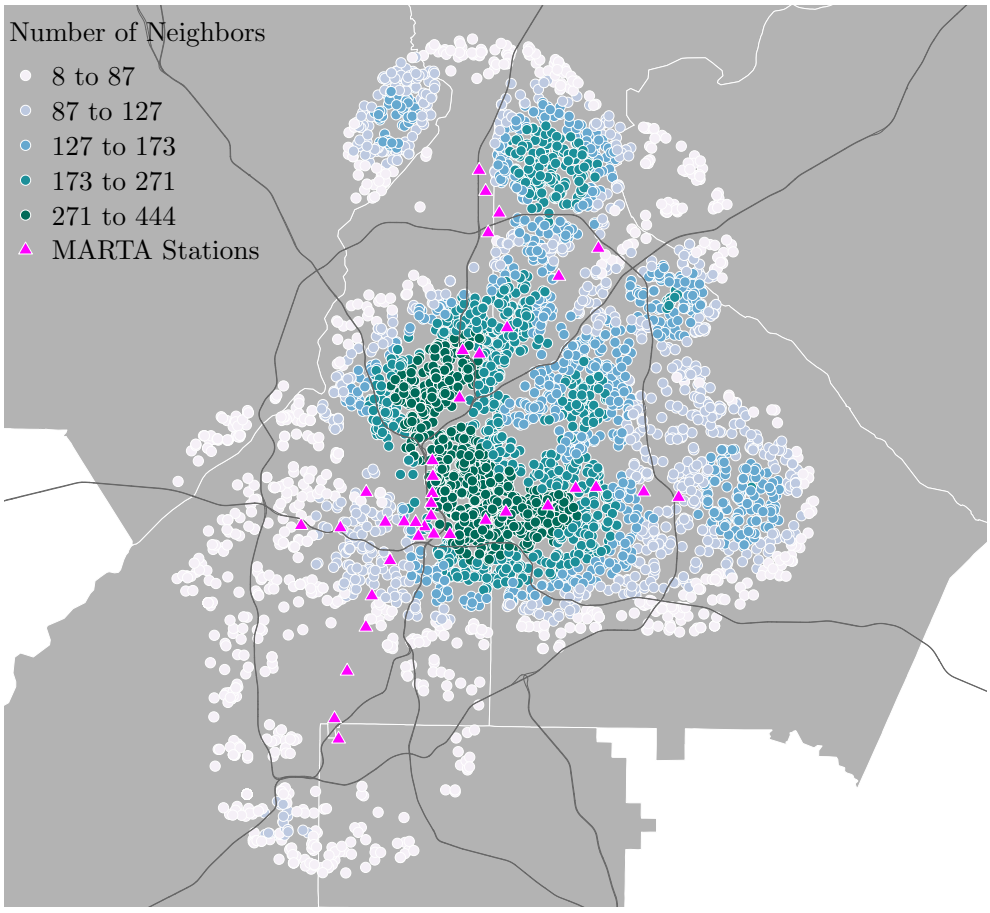

Figure 2: Observations by number of neighbors.

Table 8: Estimated model parameters and statistics.

	OLS		SAR		SEM		SDM			
Covariate	β	t-stat	β	t-stat	β	t-stat	β	t-stat	γ	t-stat
ρ			0.677^{***}	67.5			0.859^{***}	51.8		
λ					0.975^{***}	171.6				
(Intercept)	-1.886^{***}	-13.9	-3.506^{***}	-32.6	-0.959^{***}	-4.9	-1.431^{***}	-5.3		
Property type: *ref. Single unit*										
Multiple Units	-0.043^{*}	-2.1	-0.139^{***}	-8.6	-0.312^{***}	-17.2	-0.319^{***}	-17.4	0.386^{***}	10.0
Home age	0.000	1.3	-0.001^{***}	-3.5	-0.003^{***}	-14.3	-0.003^{***}	-14.8	0.005^{***}	9.0
log(Square feet)	0.915^{***}	57.0	0.731^{***}	56.1	0.783^{***}	64.5	0.776^{***}	63.7	-0.521^{***}	-12.0
log(Lot acres)	0.046^{***}	3.8	0.025^{**}	2.6	0.090^{***}	9.7	0.091^{***}	9.8	-0.139^{***}	-5.1
Race: *ref. White*										
African-American	-0.295^{***}	-21.6	-0.044^{***}	-3.9	-0.042^{***}	-3.4	-0.038^{**}	-3.1	0.021	0.9
Hispanic	-0.105^{***}	-3.3	-0.008	-0.3	-0.029	-1.4	-0.027	-1.2	0.009	0.1
log(Income)	0.263^{***}	30.5	0.065^{***}	8.8	0.075^{***}	9.5	0.070^{***}	8.9	-0.011	-0.6
log(Distance from MARTA)	-0.192^{***}	-25.2	-0.142^{***}	-23.8	-0.030^{*}	-2.2	-0.006	-0.4	-0.027	-1.4
log(Distance from freeway entrance)	-0.047^{***}	-5.4	-0.041^{***}	-5.9	0.028^{*}	2.3	$0.026^{\dagger}$	1.9	$-0.034^{\dagger}$	-1.7
N	4,917		4,917		4,917		4,917			
log($\mathcal{L}$)	-2,036		-875		-331		-271			

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 9: Lagrange multiplier tests for spatial effects.

Test	Statistic	p-value
LM_ρ	$3,208$	0
LM_λ	$8,184$	0
RLM_ρ	599	0
RLM_λ	$5,575$	0

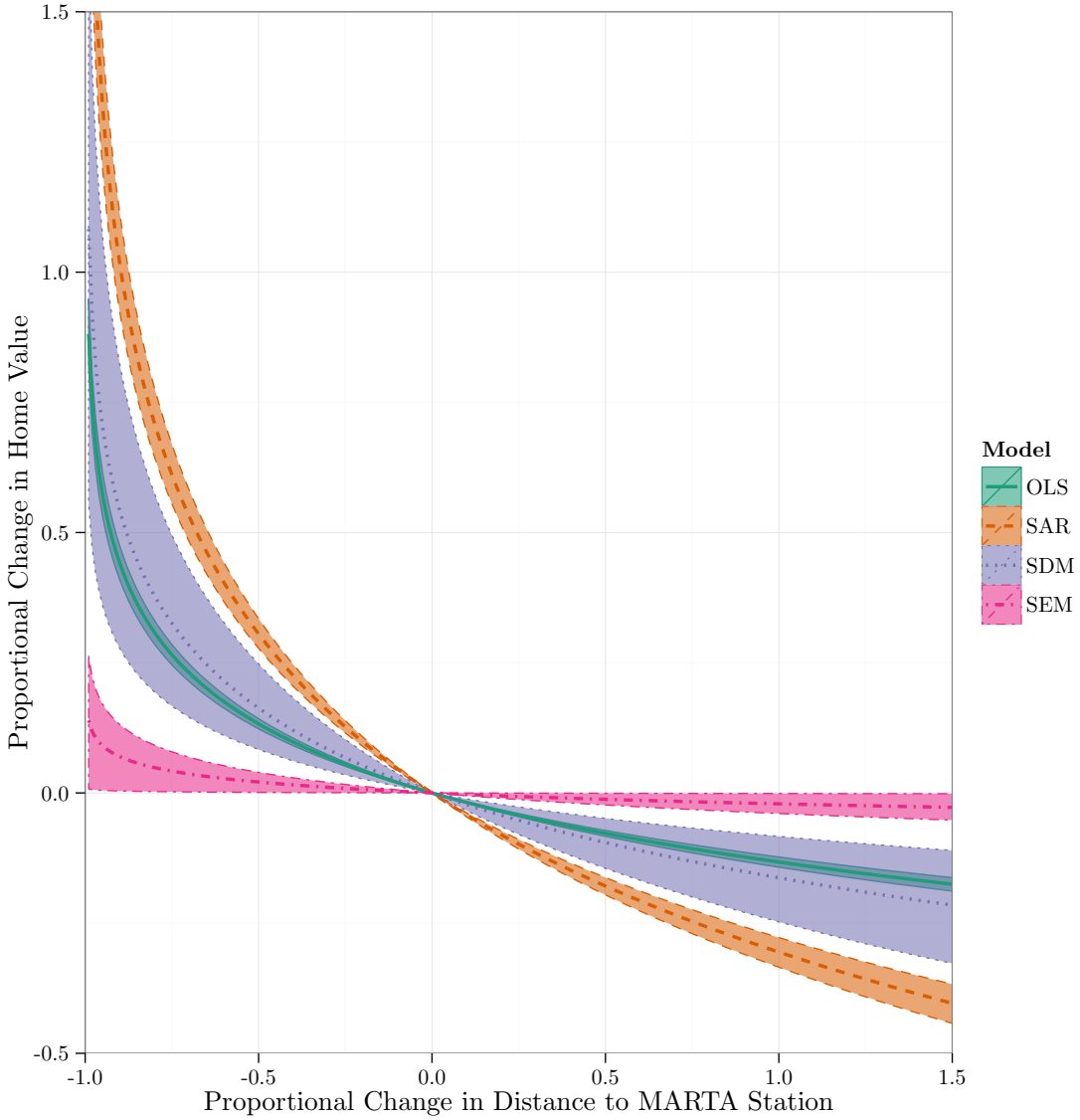

Figure 3: Estimated elasticity for transit proximity under different specifications, with confidence intervals.

Table 10: Average marginal effect of model variables.

	OLS		SAR		SEM		SDM	
Covariates	Effect	t-stat	Effect	t-stat	Effect	t-stat	Effect	t-stat
Direct Effects								
Multiple Units	−0.0435	−2.1	−0.142	−8.6	−0.312	−17.2	−0.312	−17.3
Home age	0.000307	1.27	−0.000674	−3.49	−0.00276	−14.3	−0.00273	−14.3
log(Square feet)	0.915	57	0.747	56.9	0.783	64.5	0.785	64.7
log(Lot acres)	0.0463	3.83	0.0253	2.6	0.0898	9.73	0.087	9.47
African-American	−0.295	−21.6	−0.045	−3.9	−0.0417	−3.35	−0.0389	−3.13
Hispanic	−0.105	−3.33	−0.0085	−0.337	−0.0288	−1.36	−0.0273	−1.17
log(Income)	0.263	30.5	0.066	8.86	0.0747	9.53	0.0733	9.36
log(Distance from MARTA)	−0.192	−25.2	−0.146	−23.8	−0.0301	−2.17	−0.00844	−0.568
log(Distance from freeway entrance)	−0.047	−5.37	−0.0416	−5.92	0.0283	2.32	0.0249	1.88
Indirect Effects								
Multiple Units			−0.288	−8.16			0.789	3.45
Home age			−0.00136	−3.47			0.0152	4.8
log(Square feet)			1.51	22.4			0.999	4.18
log(Lot acres)			0.0512	2.61			−0.429	−2.43
African-American			−0.0909	−4.06			−0.0799	−0.587
Hispanic			−0.017	−0.334			−0.0923	−0.137
log(Income)			0.134	9.56			0.345	3.37
log(Distance from MARTA)			−0.295	−17.2			−0.227	−3.43
log(Distance from freeway entrance)			−0.0844	−5.59			−0.0848	−1.1
Total Effects								
Multiple Units			−0.43	−8.42			0.477	2.09
Home age			−0.00204	−3.49			0.0124	3.93
log(Square feet)			2.26	31.8			1.78	7.42
log(Lot acres)			0.0765	2.61			−0.342	−1.93
African-American			−0.136	−4.02			−0.119	−0.868
Hispanic			−0.0255	−0.335			−0.12	−0.175
log(Income)			0.2	9.5			0.418	4.06
log(Distance from MARTA)			−0.441	−20.4			−0.235	−3.89
log(Distance from freeway entrance)			−0.126	−5.74			−0.0599	−0.836

The estimated effects of the model variables on home price are given in Table 10; the distribution of these effects is calculated empirically with repeated draws from the parameter variance-covariance matrices. The added information gained from modeling spatial dependence can be seen by examining the effects of home age on home price in detail. In the OLS model, a one-year increase in the age of a home cannot be said to have any relationship on its value. After controlling for correlated errors (with the SEM), an additional year is measured to have a -0.275 percent impact, a small estimate with a high degree of statistical significance. Controlling for spatial dependence (with the SAR) similarly shows negative direct, indirect, and total effects of an additional year in age. But in an SAR, the direct and indirect effects (and consequentially, the total effects) are required to have the same sign, a condition relaxed in an SDM. Indeed, the SDM shows a significant relationship between a home's age and its value but the direct and indirect effects conflict, with a direct effect of -0.003 and an indirect effect of 0.015. This is intuitive: living in an older home for its own sake brings few benefits, but living around older homes might be a sign of situation in a more established neighborhood. Similar logic can be applied to interpreting the effects of multiple unit dwellings: OLS, SAR, and SEM all suggest that the market values condominiums less than detached homes, but the SDM suggests that nearby condominiums *contribute* value. This, again, is intuitive. Isolated condominiums in neighborhoods where single-family homes are typical will be seen as less attractive. On the other hand, condominiums dominate the housing market in some of Atlanta's most exclusive neighborhoods, and nearby homes are the more valuable for it.

Our variable of primary interest, the "Miles from MARTA" variable, is significant at the 95% confidence level in each of the models, though the direct effect in the SDM is not significant (in this specification, a more negative value represents a higher MWTP). As discussed in Section 2, the total effect is the most appropriate measurement for MWTP for transit accessibility. Using the SDM model, we estimate that doubling the distance between a home and its nearest transit station lowers the expected value of the home by 23.5 percent, all else equal. Figure 3 shows 95% confidence bands of our estimated MWTP for transit proximity. As per the discussion in Section 3.2, we expect in our case that the OLS and

SEM models have biased parameters, and that the OLS and SAR models have unreliable confidence intervals. An important item of note is that the confidence intervals around the SDM model are considerably wider than either the OLS or SAR models. These confidence intervals are crucial from a risk management perspective: a conservative land value capture forecast will use not the mean MWTP, but some lower percentile so that revenues are more likely to exceed the forecast. That our mean estimate of the SDM effects is not radically different from the mean OLS effect is fortunate, but should not be generally expected.

3.5 Discussion

Autoregressive hedonic models that explicitly account for spatial dependence are not new to general econometrics and real estate science (Can and Megbolugbe, 1997; Dubin et al., 1999; Pace, 1997), and have been applied to estimate MWTP for transportation amenities. Haider and Miller (2000) showed spatial dependence was a significant issue in hedonic models of the Toronto market with respect to that city's transportation system. Armstrong and Rodríguez (2006) applied an SAR model to estimate the MWTP for access to commuter rail in the Boston suburbs, but did not comment on the effects of this dependence for their model estimates. Martínez and Viegas (2009) showed that MWTP estimates obtained with an SAR in Lisbon were similar to those obtained using OLS.

Comprehensive analyses of spatial dependence and correlation together are increasingly common. Three recent studies in particular compare multiple autoregressive structures in a transportation or accessibility context. Osland (2010) selected the SDM after deciding that the LM tests were inconclusive, thus in effect changing from the classical to general selection framework. Löchl and Axhausen (2010) selected the SEM as the most appropriate model for a hedonic forecast in Zürich's UrbanSim land use model (Waddell et al., 2003), again applying the classical framework. In this case the LM tests were conclusive, with the RLM_λ test statistic about one hundred times greater than the RLM_ρ test statistic. Ibeas et al. (2012) similarly select the SEM model to estimate MWTP for transit accessibility in Santander; these authors do not report their LM test statistics, but they reject the SDM on account of some insignificant lagged parameters. According to the general selection

Table 11: Comparison to related studies.

Statistic	Osland	Löchl and Axhausen	Ibeas et al.	This Study
$\ln(\mathcal{L}_{SDM})$	96.9	4192.6	-2.1	-272.7
$\ln(\mathcal{L}_{SEM})$	80.9	4118.0	-34.8	-338.1
$\ln(\mathcal{L}_{SAR})$	65.8		-75.0	-838.8
$\ln(\mathcal{L}_{OLS})$	52.0	3183.3	-111.9	-2006.7
$-2(\ln(\mathcal{L}_{SEM}) - \ln(\mathcal{L}_{SDM}))$	31.9*	149.2*	65.5*	130.9*
Classical	SEM**	SEM	Unknown†	SEM
General	SDM	SDM	SDM	SDM

* Reject null hypothesis with $p < 0.01$.

** Selected SDM after testing for common factors.

† Did not report LM statistics, but selected SEM.

framework, insignificant lagged parameters lead to the SAR model.

Would any of these authors have selected a different model with the general framework? Plainly, yes. The model likelihood values from each of these studies (including the present) are given in Table 11. In all four cases, the common factors test rejects that the SDM and SEM have equivalent likelihoods, and that the SDM should be the preferred model. It is therefore possible that these authors selected a model with potentially biased parameter estimates.

As mentioned in the Introduction and shown in Table 6, autocorrelation in the model residuals is a nuisance that affects the estimated standard errors of the model parameters but not the parameter estimates themselves. Autocorrelation in the dependent variable, by contrast, is a substantive problem that will bias model parameters. Selecting the SEM when a SAR or SDM is the true model may result in biased parameters (reflected in Figure 3), whereas selecting an SDM when the SEM or the SAR is the true model merely sacrifices degrees of freedom to estimate unnecessary parameters.

It is this last point that provides perhaps the greatest argument for the general framework. Standard null hypothesis significance testing is constructed to minimize the possibility of Type I error, or incorrectly rejecting a true null hypothesis. In the classical framework, the null hypothesis is that spatial effects are not present; the consequence of falsely rejecting this null hypothesis is an inefficient model. In the general framework, the consequences of

falsely rejecting the null hypothesis of spatial effects are biased parameter estimates and/or invalid parameter significance tests. Analysts should attempt to minimize the risk of selecting an inconsistent model, by beginning with the SDM and only abandoning it if spatial dependence or spatial correlation are shown to have a low probability of occurrence.

3.6 Conclusion

Accurate estimates of MWTP for public transportation infrastructure are essential for regional transportation models and plans. It is therefore imperative that analysts select an econometric structure for their models that appropriately represents the complexity of the housing market that they seek to study. Spatial effects represent both a challenge and an opportunity for such models. If spatial dependence and correlation are not considered then estimates of MWTP may be unreliable. If spatial dependence and correlation are shown to exist, on the other hand, the analyst can use these effects to develop parsimonious and powerful models.

In this paper, we have shown that considering spatial dependence and correlation in the Atlanta housing market affects estimates of MWTP for proximity to MARTA. Specifically, the estimate MWTP *less certain*, implying that a land value capture strategy built on this model should consider a substantially higher margin of error in its forecasts.

A primary contribution of this paper is its presentation and application of a model selection methodology outside of the classical framework. The literature defining spatial effects and models to accommodate them is sufficiently mature that analysts studying housing markets should assume *a priori* that these effects are present, and the burden of proof should be on their absence rather their existence. The re-orientation towards a general-to-specific framework will prevent analysts from incorrectly rejecting the conservative and general SDM in favor of a more efficient but potentially inappropriate specification.

3.7 References

Alonso, W., 1960. A theory of the urban land market. *Papers in Regional Science* 6, 149–157.

Anselin, L., 1980. Estimation methods for spatial autoregressive structures. *Regional Science Dissertation & Monograph Series, Program in Urban and Regional Studies, Cornell University* .

Anselin, L., 1988a. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis* 20, 1–17.

Anselin, L., 1988b. *Spatial Econometrics: Methods and Models (Studies in Operational Regional Science)*. Kluwer, Dordrecht.

Anselin, L., 2003. Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review* 26, 153–166.

Anselin, L., Bera, A.K., Florax, R.J.G.M., Yoon, M.J., 1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics* 26, 77–104.

Armstrong, R.J., Rodríguez, D.A., 2006. An evaluation of the accessibility benefits of commuter rail in eastern Massachusetts using spatial hedonic price functions. *Transportation* 33, 21–43.

Atlanta Regional Commission, 2012. ARC GIS data and maps. URL: `http://www.atlantaregional.com/info-center/gis-data-maps`.

Bivand, R., 2006. Implementing spatial data analysis software tools in R. *Geographical Analysis* 38, 23–40.

Bowes, D.R., Ihlanfeldt, K.R., 2001. Identifying the impacts of rail transit stations on residential property values. *Journal of Urban Economics* 50, 1–25.

Brigham, E.F., 1965. The determinants of residential land values. *Land Economics* 41, 325–334.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1999. Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science* 39, 497–524.

Burridge, P., 1981. Testing for a common factor in a spatial autoregression model. *Environment and Planning A* 13, 795–800.

Can, A., Megbolugbe, I., 1997. Spatial dependence and house price index construction. *Journal of Real Estate Finance and Economics* 14, 203–222.

Chen, H., Ruffalo, A., Dueker, K.J., 1998. Measuring the impact of light rail systems on single-family home values: a hedonic approach with geographic information system application. *Transportation Research Record* 1617, 38–43.

Chernobai, E., Reibel, M., Carney, M., 2009. Nonlinear spatial and temporal effects of highway construction on house prices. *The Journal of Real Estate Finance and Economics* 42, 348–370.

Cliff, A.D., Ord, J.K., 1970. Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography* 46, 269–292.

Debrezion, G., Pels, E., Rietveld, P., 2007. The impact of railway stations on residential and commercial property value: a meta-analysis. *The Journal of Real Estate Finance and Economics* 35, 161–180.

Dubin, R.A., 1998. Spatial autocorrelation: a primer. *Journal of Housing Economics* 7, 304–327.

Dubin, R.A., Pace, R.K., Thibodeau, T.G., 1999. Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature* 7, 79–95.

Dubin, R.A., Sung, C.H., 1987. Spatial variation in the price of housing: rent gradients in non-monocentric cities. *Urban Studies* 24, 193–204.

Florax, R.J.G.M., Folmer, H., Rey, S.J., 2003. Specification searches in spatial econometrics: The relevance of Hendry's methodology. *Regional Science and Urban Economics* 33, 557–579.

Grass, R.G., 1992. The estimation of residential property values around transit station sites in Washington, D.C. *Journal of Economics & Finance* 16, 139–146.

Haider, M., Miller, E.J., 2000. Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. *Transportation Research Record* 1722, 1–8.

Hendry, D.F., 1979. Predictive failure and econometric modelling in macroeconomics: the transactions demand for money. *Economic Modelling* , 217–242.

Hess, D.B., Almeida, T.M., 2007. Impact of proximity to light rail rapid transit on station-area property values in Buffalo, New York. *Urban Studies* 44, 1041–1068.

Iacono, M., Levinson, D., 2011. Location, regional accessibility, and price effects: evidence from home sales in Hennepin County, Minnesota. *Transportation Research Record* 2245, 87–94.

Ibeas, A., Cordera, R., Dell'Olio, L., Coppola, P., Dominguez, A., 2012. Modelling transport and real-estate values interactions in urban systems. *Journal of Transport Geography* 24, 370–382.

Kim, J., Goldsmith, P., 2008. A spatial hedonic approach to assess the impact of swine production on residential property values. *Environmental and Resource Economics* 42, 509–534.

Kressner, J.D., Garrow, L.A., 2012. Lifestyle segmentation variables as predictors of home-based trips for Atlanta, Georgia, airport. *Transportation Research Record* 2266, 20–30.

Larch, M., Walde, J., 2008. Lag or error? Detecting the nature of spatial correlation, in: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (Eds.), Data Analysis, Machine Learning and Applications, Springer, Berlin, Freiburg. pp. 301–308.

LeSage, J.P., Fischer, M.M., 2008. Spatial growth regressions: model specification, estimation, and interpretation. *Spatial Economic Analysis* 3, 275–304.

LeSage, J.P., Pace, R.K., 2009. *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.

Lewis-Workman, S., Brod, D., 1997. Measuring the neighborhood benefits of rail transit accessibility. *Transportation Research Record* 1576, 147–153.

Löchl, M., Axhausen, K.W., 2010. Modeling hedonic residential rents for land use and transport simulation while considering spatial effects. *Journal of Transport and Land Use* 3, 39–63.

Martínez, L.M., Viegas, J.M., 2009. Effects of transportation accessibility on residential property values. *Transportation Research Record* 2115, 127–137.

Massell, B.F., Stewart, J.M., 1971. The determinants of residential property values. *Institute for Public Policy Analysis, Stanford University, Discussion Paper* .

McMillen, D.P., McDonald, J., 2004. Reaction of house prices to a new rapid transit line: Chicago's Midway Line, 1983-1999. *Real Estate Economics* 32, 463–486.

Mikelbank, B.A., 2004. Spatial analysis of the relationship between housing values and investments in transportation infrastructure. *The Annals of Regional Science* 38, 705–726.

Mur, J., Angulo, A.M., 2006. The spatial Durbin model and the common factor tests. *Spatial Economic Analysis* 1, 207–226.

Nelson, A., 1992. Effects of elevated heavy-rail transit stations on house prices with respect to neighborhood income. *Transportation Research Record* 1359, 127–132.

Osland, L., 2010. An application of spatial econometrics in relation to hedonic house price modeling. *The Journal of Real Estate Research* 32, 289–320.

Pace, R.K., 1997. Performing large spatial regressions and autoregressions. *Economics Letters* 54, 283–291.

R Development Core Team, 2013. R: A Language and Environment for Statistical Computing. URL: `http://www.r-project.org`.

Ridker, R.G., Henning, J.A., 1967. The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics* 49, 246–257.

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *The Journal of Political Economy* 82, 34–55.

Smith, J.J., Gihring, T.A., 2006. Financing transit systems through value capture. *American Journal of Economics and Sociology* 65, 751–786.

Tax Policy Center, 2013. What are the sources of revenue for local governments? URL: `http://www.taxpolicycenter.org/briefing-book/`.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46, 234–240.

U.S. Census Bureau, 2010. Wealth and Asset Ownership. URL: `http://www.census.gov/people/wealth/`.

Waddell, P.A., Borning, A., Noth, M., Freier, N., Becke, M., Ulfarsson, G., 2003. Microsimulation of urban development and location choices: design and implentation of UrbanSim. *Networks and Spatial Economics* 3, 43–67.

CHAPTER IV

TRANSIT INFRASTRUCTURE AND HOME PRICE STABILITY

Gregory S. Macfarlane and Juan Moreno-Cruz

Working Paper, 2014

Chapter Abstract

Public transit infrastructure and other features of the urban environment shape housing markets, as neighborhoods with high accessibility also tend to be highly valued. But housing markets are dynamic and sometimes turbulent, and the role that transit infrastructure plays in long-term home value in the face of macroscopic events has not been studied. In this paper, we model the performance of the Atlanta housing market in the period from 2002-2012 as a function of a home's proximity to the MARTA rail network. Univariate spatial Durbin models show homes in proximity to public transit had higher values, higher growth rates, and less volatility in growth over the period than homes further away. Multivariate latent class mixture models confirm these results and also show that the Atlanta housing market can be considered as two distinct classes: homes near to MARTA are more likely to be in a class with positive value growth over the period.

4.1 Background

The aggregate story of the US housing market in the period from 2000 through 2012 can be told in three parts, illustrated by the plot of the Case-Shiller home price index in Figure 4 (Standard & Poors, 2013). From 2000 through 2005, home prices rose sharply, with the average home selling in late 2005 for almost twice its 2000 value. After leveling off through 2006, home prices fell precipitously through 2007 and 2008. The market has remained depressed in the intervening years, though a turning point may have been reached in the

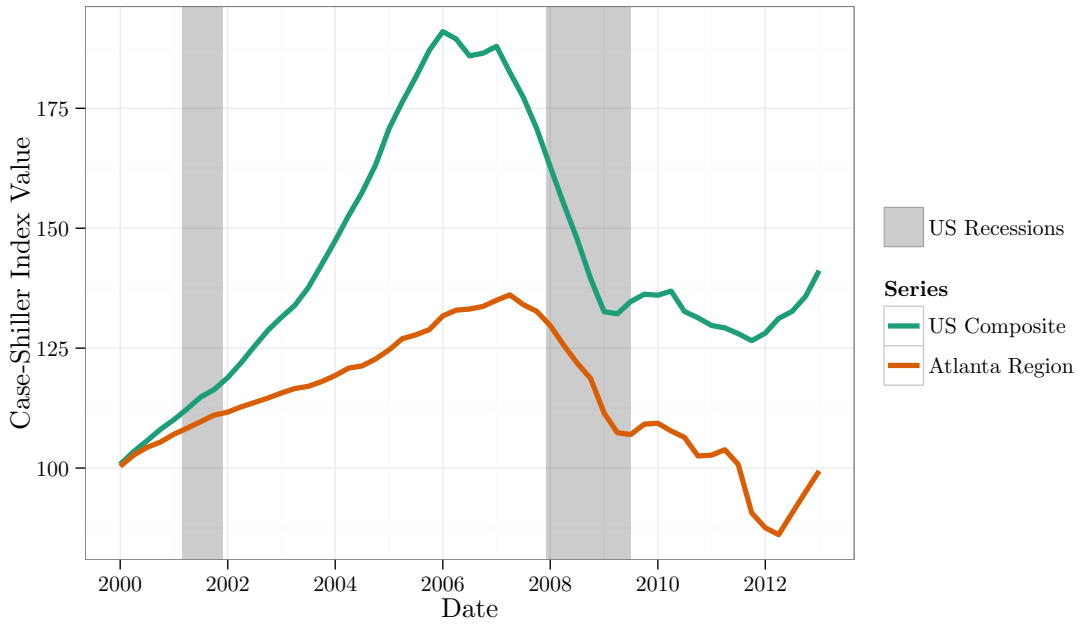

Figure 4: Case-Shiller home price index, seasonally adjusted.

winter of 2011-2012. But the story told by the composite Case-Shiller index is not homogeneous across the US. Some markets outperformed the national average, gaining more in value during the boom and losing less during the bust. Others underperformed. In the Atlanta market specifically (also shown in Figure 4), the 2007 apex was substantially lower than the national average, and the nadir was more severe. What these aggregate indices do not show, however, is that the heterogeneity in the response goes deeper. Homes in some neighborhoods remained stable in value throughout the period, others fluctuated wildly, and some others may even have gained value in the first five years of the new century without subsequently losing it.

This heterogeneity in price performance is not necessarily surprising. Neighborhoods are themselves heterogeneous; every neighborhood has its own unique blend of amenities. Homes after all are consumable goods as much as they are financial investments, and households will pay for the amenities that they value; it is possible that the relative value of these amenities changes with (and perhaps because of) the overall economy. But understanding which neighborhood characteristics correlate with advantageous price performance — and which of these characteristics planners can influence — is an important consideration for

urban policy.

One possible amenity is transportation accessibility, and access to public transit in particular. Neighborhoods with access to public transportation have a higher value, as identified by both theory (Alonso, 1960) and empirical observation (Lewis-Workman and Brod, 1997; Iacono and Levinson, 2011). There are also several mechanisms by which public transit accessibility may be associated with better long-term value performance. It may be that transit-supportive development (such as high population densities) constrains the housing supply, preventing overbuilding. It may also be that homes with better access to a region's opportunities are able to adapt more quickly to economic changes; laid-off workers in such homes may be able to find a new job more quickly and avoid relocation or default (as was observed by Pivo (2013)).

This study presents an empirical investigation of home price growth and volatility in Fulton County, Georgia between 2002 and 2012, testing the theory that public rail transit infrastructure — specifically, the Metropolitan Atlanta Rapid Transit Authority (MARTA) rail network — is correlated with positive growth or stability in a housing market. Spatial Durbin models of average value, average growth, and variance in growth show that neighborhoods closer to MARTA stations have higher average values, higher average growth rates, but higher growth volatility than homes further away. These findings are based on terms interacted with the growth in neighborhood income. Further, latent class mixture models reveal at least two distinct home markets in the Atlanta area; homes with access to the MARTA network[1] are more likely to belong to a class of homes that experienced positive price growth over the period in question.

The paper proceeds as follows. The rest of this section discusses the relevant literature and places the study a theoretical context. Section 4.2 describes the dataset used in this analysis, which is constructed from the Fulton County tax assessor's database and other publicly-available sources. Section 4.3 presents univariate spatial regressions on value, growth, and volatility; Section 4.4 presents a multivariate latent class mixture model to investigate the relationship between transit proximity and potential submarkets in the study

[1] Access to the network is provided by proximity to its stations.

area. The study concludes with an interpretation of the findings and an outline for further investigation.

4.1.1 Literature

Numerous authors have identified a correlation between investment in public transit and increased home values. These studies can generally be classified into two broad types. The first type of study considers home prices in a particular period and relates the market's willingness-to-pay for transit proximity (e.g., Lewis-Workman and Brod, 1997; Bowes and Ihlanfeldt, 2001; Debrezion et al., 2007); these studies of necessity assume that the housing market is in some sort of equilibrium. The second study type measures price changes in response to transit construction, assuming that unobserved or endogenous market characteristics can be differenced out (e.g., Grass, 1992; McMillen and McDonald, 2004). Both types of study have generally shown homes near transit stations are more valuable in equilibrium and expanding transit infrastructure results in increased home values. The consequences of these findings are twofold: it may be possible for governments to recoup the cost of investment through elevated property taxes (Smith and Gihring, 2006); however, elevated property values and resulting gentrification may displace the very populations who most rely on public transit (Pollack et al., 2010).

In the first type of study the housing market and transit infrastructure are both fixed. In the second type, the housing market is considered fixed[2] as transit infrastructure changes. The existing literature — to the best of our knowledge — is missing a potential third type of study, where the transit network remains constant *as the housing market changes*. In particular, we ask, "are homes in proximity to transit networks *more resilient* to demand-related shocks to the housing market?"

There are economic theories that explain why differentials between and within housing markets exist, and how public transportation infrastructure or other features of the built environment may contribute to this heterogeneity. Glaeser et al. (2008) presents a model that predicts inelastic housing markets — those with constraints on construction — will

[2]With the exception of possible time or neighborhood-level fixed effects.

perform better during an exogenous and irrational home price bubble. Specifically, inelastic markets cannot build new homes to meet an imagined demand, and so supply is not artificially inflated. Prices rise in response to demand, but return to their proper value when demand subsides. In elastic markets by contrast, the homes' supply expands and keeps prices from rising but results in an oversupply and a consequent price collapse when the perceived demand returns to its real level. Home prices in Atlanta rose more modestly than the national average from 2000-2006, and dropped below their 2000 level by mid-2011. The theory suggests that, on average, the Atlanta housing market is quite elastic; but the Atlanta market might actually contain multiple submarkets. Considering Fulton County in particular, we might hypothetically identify at least two housing markets: an elastic market in the suburban north and southwest, and an inelastic market in the inner city. In this case, proximity to MARTA rail may serve as an indicator of an urban environment and presumably constrained development. Homes in neighborhoods close to MARTA would show higher values but potentially *more* volatility under this model.

Guerrieri et al. (2013) presents another explanatory model of uneven response to demand changes. This spatial model considers the location of wealthy and poor residents in response to an exogenous increase in demand. The model predicts that wealthy residents displace poor residents in neighborhoods adjacent to existing wealthy neighborhoods. Home values in these gentrifying neighborhoods would therefore increase in value more quickly than the rest of the market during the demand shock. In the context of this study, the Guerrieri et al. results suggest that neighborhood gentrification may be a confounding variable: advantageous home price performance in neighborhoods close to MARTA stations could be a manifestation of rising incomes in that neighborhood rather than an externality of transit proximity. On the other hand, neighborhoods near MARTA stations could be likely candidates for gentrification precisely *because* MARTA is nearby. We control for this gentrification effect and identify an interesting new insight that suggests an interaction between gentrified neighborhoods and proximity to MARTA rail.

4.2 Data

Data for this study was supplied by the Fulton County, Georgia tax assessor's office in response to a public records request. The appraised or assessed value of every single-unit residential property in the county from 2002 through 2012 is available, as is basic information about the structure such as the year of construction, the number of rooms in the building, and the size of the property. Of particular note is the "effective age," which is the age of the structure discounted by the tax assessor to accommodate reconstructions, renovations, and installed amenities that may not be original to the home (e.g., air conditioning, indoor plumbing, etc.). Only home values from Fulton County are used to avoid idiosyncratic appraisal methods between counties and to ensure that a common set of covariates is available.

Several additional variables have been appended to these records. As a measure of neighborhood wealth, we use the home's Census tract median income recorded in the 2010 American Community Survey (U.S. Census Bureau, 2010b). Similarly, we measure the neighborhood racial composition as the percent of white people residing in each Census tract (U.S. Census Bureau, 2010a). We develop a measure for gentrification as the percent increase in Census tract real median income from 2000 to 2010. To measure proximity to public transportation, we measure the Euclidean distance[3] from each home to the nearest MARTA station. We also measure the Euclidean distance to the nearest freeway entrance point, as an indicator of highway accessibility.

Although we have price evaluations available for each home in each of the ten years from 2002-2013, panel regression methods are inadequate, as the independent variables do not change for approximately 99% of homes over the time period in question. In particular, the MARTA rail network remained constant for all observations. It is therefore necessary to develop metrics by which the price performance of the home over time can be characterized.

[3]Our explorations found no meaningful difference between Euclidean and network distance.

4.2.1 Price Performance

Good investments express a number of characteristics. First, they have a positive net value: in a given period t, the expected value V of an investment i should be higher than its initial value,

$$\mathrm{E}\left[V_i(t)\right] > V_i(0) \tag{29}$$

with better investments having higher expected values. Additionally, good investments should have positive growth: the average change in value is positive,

$$\mathrm{E}\left[\Delta V_i(t)\right] = \mathrm{E}\left[\frac{V_i(t+1) - V_i(t)}{V_i(t)}\right] > 0 \tag{30}$$

with better investments having higher expected growth rates. Finally, good investments have predictable growth: the variation in the growth rate is minimized,

$$\mathrm{Var}\left[\Delta V_i(t)\right] = 0. \tag{31}$$

An abstract investment's performance might be generalized with the following function:

$$V(t) = e^{rt} + \alpha e^{at}\sin(\omega t) \tag{32}$$

Where $V(t)$ is a function of the exponential growth rate r, the volatility α, the rate at which the volatility is increasing a, and the frequency of oscillation ω. Excellent investments will have high r and $\alpha, a, \omega = 0$, thus reducing Equation 32 to the elementary long-term growth equation, e^{rt}. Figure 5 shows a collection of eight value paths created by changing r, α, a, and ω, in addition to the Case-Shiller index for the Atlanta market. A table showing the average value, the average growth rate, and the standard deviation of the growth rate for each function is given in Table 12. These three indicators sufficiently discriminate between the functions, as well as characterize the Case-Shiller index as being generally positive but with negligible average growth and high growth rate volatility.

These three metrics — mean value, mean growth rate, and standard deviation of the growth rate — were calculated for each of the homes in the dataset. The three functions used indexed values $V(t)/V(0), t \in 0, \ldots, 9$ to cancel the effect of initial price. Full descriptive statistics of all variables are given in Table 13.

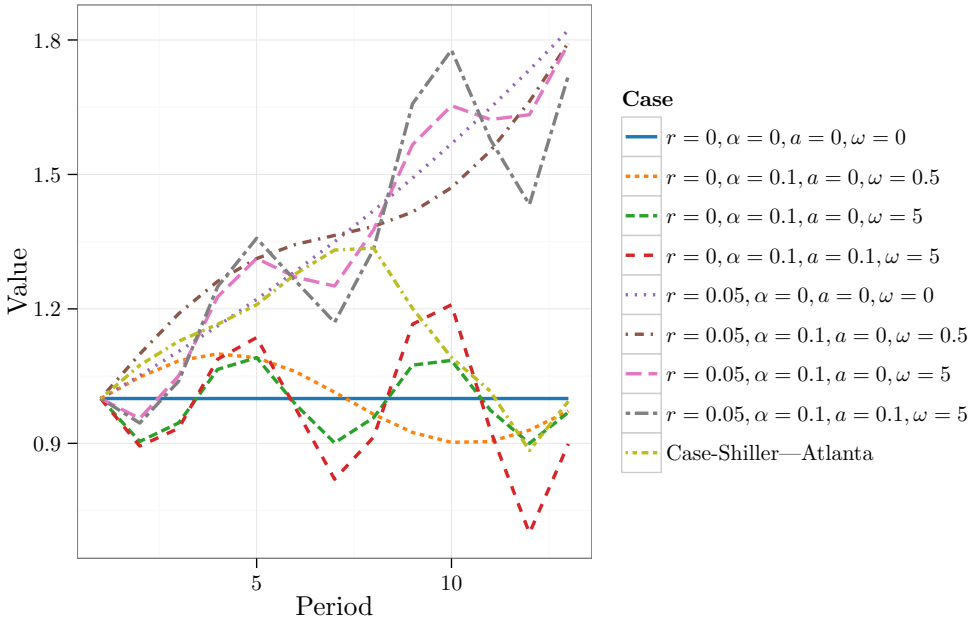

Figure 5: Illustrative cases of an abstract investment's value performance.

Table 12: Volatility metrics for example cases.

	Value	Growth Rate	
Value Function	Mean	Mean	Std. Dev.
Case-Shiller (Atlanta)	1.13	0.003	0.08
$r = 0.05, \alpha = 0, a = 0, \omega = 0$	1.37	0.051	0
$r = 0.05, \alpha = 0.1, a = 0, \omega = 5$	1.36	0.052	0.071
$r = 0.05, \alpha = 0.1, a = 0, \omega = 0.5$	1.37	0.05	0.029
$r = 0.05, \alpha = 0.1, a = 0.1, \omega = 5$	1.35	0.054	0.129
$r = 0, \alpha = 0, a = 0, \omega = 0$	1	0	0
$r = 0, \alpha = 0.1, a = 0, \omega = 5$	0.989	0.001	0.088
$r = 0, \alpha = 0.1, a = 0, \omega = 0.5$	1	-0.002	0.036
$r = 0, \alpha = 0.1, a = 0.1, \omega = 5$	0.974	0.007	0.184

Table 13: Descriptive statistics of model variables

	Notes	Mean	Median	Std. Dev.	1%	99%
Independent variables						
Property acreage		0.544	0.289	2.6	0.015	4.58
Number of rooms		6.69	6	1.89	3	12
Effective age	Appraised, considers age of structure and other amenities.	35.7	30	19.9	11	93
Median income	2010 ACS Census tract level.	74,435	67,864	40,790	15,893	196,875
Percent white	Percent of white population in Census tract, 2010 ACS.	0.514	0.649	0.352	0	0.954
Income growth	Percent increase in Census tract income, 2000-2010, real dollars.	0.211	0.157	0.324	-0.311	1.41
Distance to MARTA station (miles)		4.65	3.1	4.1	0.204	15.1
Distance to freeway entrance		1.82	1.34	1.46	0.183	6.41
Dependent variables						
Mean value		1.05	1.03	0.232	0.79	1.77
Mean growth rate		-0.023	-0.011	0.06	-0.13	0.125
Std. deviation of growth rate		0.134	0.105	0.135	0.013	0.54

4.3 Spatial Analysis

We wish to describe a home's value performance as a function of its attributes, including proximity to a MARTA rail station. Following on the presentation in Chapter 3, it is necessary to control for spatial dependence, correlation, and endogenous missing variables. We do this with a *spatial Durbin model* (SDM),

$$\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{33}$$

where $\boldsymbol{y}$ may be one of the three characterization functions. Other model elements include X, an $n \times p$ matrix where n is the number of homes and p is the number covariate attributes; $\boldsymbol{\epsilon}$, a stochastic error component assumed to have an independent and identical normal distribution; and W, an $n \times n$ matrix of weights mapping the spatial relationship between all i, j pairs $\{i, j\} \in 1, \ldots, n$. The model parameters $\rho, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are estimated by maximum likelihood (ML). The average marginal direct, indirect, and total effects of a covariate $\boldsymbol{x}_k, k \in 1, \ldots, p$ in the SDM are

$$\begin{aligned} M(k)_{\text{direct}} &= n^{-1} tr((I - \rho W)^{-1}(I\beta_k + W\gamma_k)) \\ M(k)_{\text{total}} &= n^{-1}\boldsymbol{\iota}'(I - \rho W)^{-1}(I\beta_k + W\gamma_k)\boldsymbol{\iota} \\ M(k)_{\text{indirect}} &= M(k)_{\text{total}} - M(k)_{\text{direct}} \end{aligned} \tag{34}$$

where $\boldsymbol{\iota}$ is a vector of ones of length n. These effects and their empirical t-statistics can be obtained through a Monte Carlo simulation. We estimate model parameters and effects using maximum likelihood (ML) routines included in the `spdep` package for R (Bivand, 2013). It should be noted that the log-likelihood function for the SDM includes a log-determinant term $\ln|I - \rho W|$ that must be evaluated at every iteration of the maximization algorithm. This is a computationally expensive process of an order increasing with n. For this reason, we restrict our analysis to a random sample of $n = 5000$ observations.

In this study, we use a weights matrix W that considers the 50 nearest observations as neighbors, with a link weighted by the inverse distance between the observations. This weighting scheme helps to accommodate the changing density across the study area, and

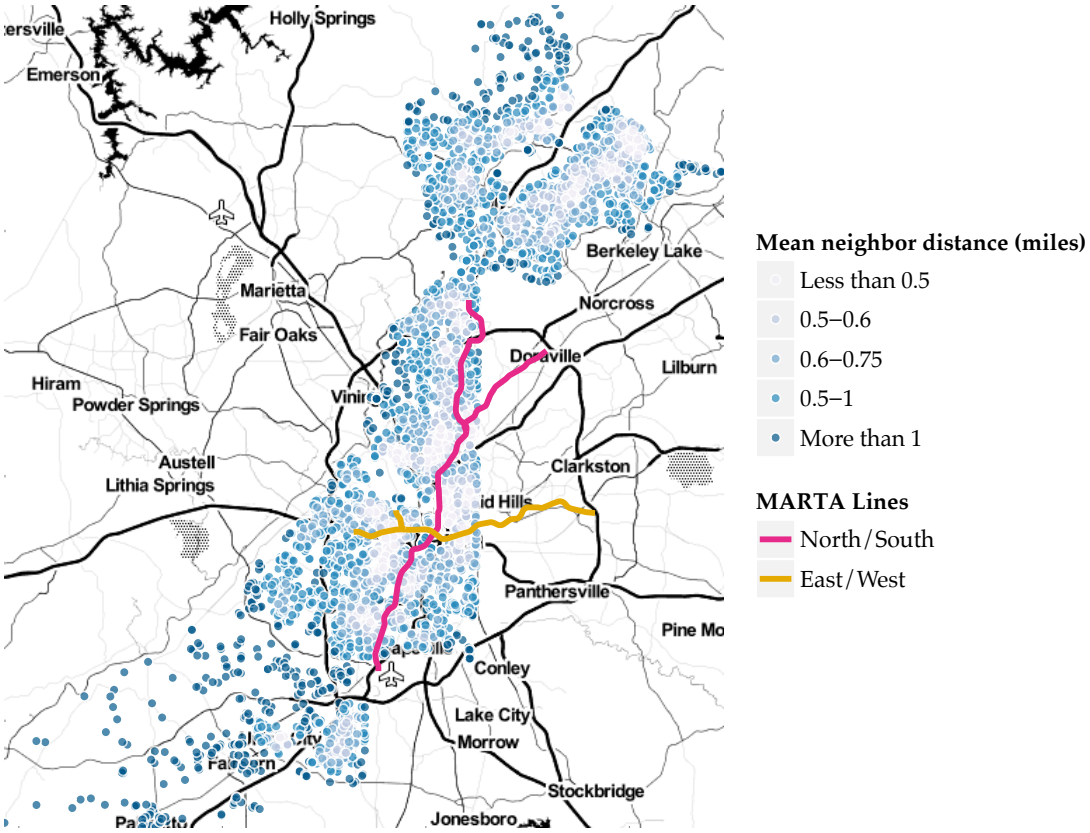

Figure 6: Observations in spatial models, by average distance of 50 nearest neighbors.

was found to have a good fit in the analysis described in Appendix B.[4] Figure 6 shows the neighborhood density in W by comparing the average distance between a home and all of its neighbors throughout the study area. Homes near to the region's core (and to other cities in the north part of the county) have a very low average neighbor distance, compared to suburban and exurban homes.

In all of our models, we use a log-log specification, applying logarithmic transformations to each $\boldsymbol{y}$ and to the $\boldsymbol{x}$ variables representing property size, home age, median income, distance to rail, and distance to a freeway entrance. In this case, the estimated parameters express a constant elasticity, aiding interpretation; also, the ML algorithm converges more easily if the parameter estimates are on the same scale.

[4]Even though the weights matrices developed in the appendix were built for other data, the W is meant to be specific to the region, and not the data.

4.3.1 Results

We estimated SDM models of mean value, growth rate, and growth volatility with three sets of covariates. This results in a total of nine models. The simulated effects for each model and associated significance statistics are given in Tables 14 and 15; whereas Table 14 presents all three effects types for the mean value models, Table 15 shows only the total effects for the growth rate and volatility models.

Beginning with the mean value models in Table 14, we can identify many of the trends previously observed in Chapter 3. Take the effective age, for instance: as the age of a home increases, its expected mean value decreases, but it goes up with the age of its neighbors. As before, it is best to own a new home in an old (and presumably established or elite) neighborhood. The racial composition of a home's Census tract has no distinguishable direct effect, but rather a strong indirect effect with the expected value increasing with the share of white households. As the income of a home's neighbors increases, the expected value of a home *decreases*. This initially unintuitive result could be explained first by the fact that the absolute value of a home has been normalized out of the equation, and second by the observations of Anderson and Beracha (2010) that home prices in wealthier neighborhoods are more sensitive to fluctuations in capital markets. Perhaps wealthier people lost more of their assets in the financial crisis, reducing demand (and subsequently prices) in wealthier neighborhoods.

The effect of interest is the log(Distance to Rail) parameter, and in particular the total effect. In the "Access" model the estimated parameter of -0.02 is weakly significant, but has the hypothesized sign: all else equal, a home's average value decreases as the distance between the home and the rail station increases. The "Gentrify" model explores the possibility that the observed effect is not due to transit proximity *per se*, but rather neighborhood gentrification. Indeed, neighborhood income growth over the period is strongly correlated with an increase in home values, consistent with what would be expected through gentrification. The final model, "Interaction," includes an interaction term for the combined effects of income growth and rail proximity. This final model rejects the Gentrify model in a likelihood ratio test (p-value 0.002). In this model, the effect of income growth on its own

Table 14: Effects of covariates on mean home value.

	Access		Gentrify		Interaction	
Covariates	$M(k)$	t-stat	$M(k)$	t-stat	$M(k)$	t-stat
Direct Effects						
log(Acres)	0.037^{***}	13.6	0.036^{***}	13.1	0.036^{***}	12.5
Total Rooms	-0.002	-1.16	-0.001	-0.806	-0.001	-0.788
log(Effective Age)	-0.063^{***}	-15.8	-0.063^{***}	-15.4	-0.063^{***}	-15.3
log(Income)	0.01	0.703	-0.017	-1.03	-0.016	-0.97
Percent White	-0.011	-0.235	0	-0.007	-0.003	-0.062
Income Growth			0.039^{**}	2.61	-0.051	-0.47
log(Distance to Rail)	0.035^{*}	2.43	$0.03^{\dagger}$	1.92	$0.028^{\dagger}$	1.82
log(Distance to Freeway)	-0.002	-0.147	-0.001	2.61	-0.002	-0.157
log(Rail Dist.)$*$ Income Growth					0.01	0.85
Indirect Effects						
log(Acres)	0.002	0.156	-0.01	-0.67	-0.005	-0.34
Total Rooms	0.004	0.37	0.014	1.24	$0.019^{\dagger}$	1.7
log(Effective Age)	0.073^{*}	2.35	0.095^{**}	3.24	0.118^{***}	4.07
log(Income)	-0.108^{**}	-2.61	-0.102^{**}	-2.61	-0.111^{**}	-2.73
Percent White	0.336^{***}	5.03	0.324^{***}	5.23	0.341^{***}	5.46
Income Growth	0.069	1.61	1.11^{***}	3.57		
log(Distance to Rail)	-0.055^{**}	-2.78	$-0.038^{\dagger}$	-1.84	-0.013	-0.621
log(Distance to Freeway)	0.029	1.28	0.032	1.61	$0.035^{\dagger}$	1.67
log(Rail Dist.)$*$ Income Growth					-0.112^{***}	-3.37
Total Effects						
log(Acres)	0.039^{*}	2.54	$0.026^{\dagger}$	1.81	0.031^{*}	2.2
Total Rooms	0.003	0.233	0.013	1.13	0.018	1.6
log(Effective Age)	0.01	0.311	0.032	1.09	$0.055^{\dagger}$	1.9
log(Income)	-0.098^{**}	-2.76	-0.118^{***}	-3.62	-0.127^{***}	-3.78
Percent White	0.326^{***}	7.11	0.324^{***}	7.7	0.338^{***}	8.36
Income Growth			0.108^{**}	2.95	1.06^{***}	4
log(Distance to Rail)	$-0.02^{\dagger}$	-1.73	-0.008	-0.725	0.015	1.23
log(Distance to Freeway)	$0.027^{\dagger}$	1.66	0.031^{*}	2.95	0.033^{*}	2.26
log(Rail Dist.)$*$ Income Growth					-0.102^{***}	-3.63
Model Log-likelihood	$3,205$		$3,214$		$3,221$	

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 15: Total effects of covariates on home performance.

	Access		Gentrify		Interaction	
Covariates	$M(k)_{\text{Total}}$	t-stat	$M(k)_{\text{Total}}$	t-stat	$M(k)_{\text{Total}}$	t-stat
Mean Growth Rate						
log(Acres)	0.013^{*}	2.32	$0.01^{\dagger}$	1.79	0.012^{*}	2.31
Total Rooms	0	−0.057	0.002	0.381	0.003	0.691
log(Effective Age)	−0.008	−0.655	−0.003	−0.297	0.004	0.355
log(Income)	−0.006	−0.497	−0.01	−0.827	−0.013	−1.07
Percent White	0.116^{***}	7.29	0.116^{***}	7.68	0.121^{***}	8.17
Income Growth			0.021	1.56	0.337^{***}	3.43
log(Distance to Rail)	$-0.007^{\dagger}$	−1.78	−0.005	−1.23	0.003	0.63
log(Distance to Freeway)	0.006	1.03	0.007	1.56	0.008	1.31
log(Rail Dist.)∗ Income Growth					-0.034^{**}	−3.24
Model Log-likelihood	8,621		8,639		8,677	
Standard Deviation of Growth Rate						
log(Acres)	$0.013^{\dagger}$	1.82	0.009	1.26	0.009	1.18
Total Rooms	−0.009	−1.62	−0.007	−1.16	−0.007	−1.17
log(Effective Age)	0.037^{*}	2.5	0.043^{**}	2.79	0.041^{**}	2.6
log(Income)	-0.051^{**}	−2.98	-0.054^{**}	−3.1	-0.055^{**}	−3.13
Percent White	-0.043^{*}	−1.99	-0.044^{*}	−2.05	-0.044^{*}	−2.03
Income Growth			0.023	1.24	−0.047	−0.339
log(Distance to Rail)	-0.012^{*}	−2.26	$-0.01^{\dagger}$	−1.72	−0.011	−1.64
log(Distance to Freeway)	$0.015^{\dagger}$	1.92	0.016^{*}	1.99	0.016^{*}	2.11
log(Rail Dist.)∗ Income Growth					0.008	0.503
Model Log-likelihood	6,180		6,183		6,184	

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

becomes more pronounced, and the interaction term is strongly significant with a value of −0.102. Given two identical homes located in identical neighborhoods that are both growing in median income, the home located closer to transit will have a higher expected value across the study period.

Table 15 presents the total effects for the three model specifications with the other dependent variables: the mean and the standard deviation of the growth rate over the period. Speaking first of the mean growth rate model, very few of the covariates are significant and show little change across the specifications. However, the interaction of income growth and rail transit proximity is significant, and as before, shows the expected sign. As income

growth in a neighborhood increases, the average growth rate of home prices in that neighborhood will rise; as the distance between those homes and a MARTA station decreases, the growth rate will rise further.

In the case of the models estimating the relationship between home and neighborhood attributes and the standard deviation of the growth rate, the interaction between income growth and MARTA proximity does not significantly improve model likelihood. The effect of MARTA proximity on price volatility is weakly significant, though its negative sign indicates that home prices closer to MARTA stations are actually *more* volatile.

4.4 Latent Class Analysis

While the previous univariate analysis is informative, home value performance is by nature a multivariate problem. We can examine the joint performance of mean value, mean growth rate, and growth rate volatility with a multivariate finite mixture model,

$$H(Y|X, w, \psi) = \sum_{k=1}^{K} \pi_k(w, \alpha_k) \prod_{d=1}^{D} f_{kd}(Y_d|X_d, \theta_{kd}) \tag{35}$$

where the mixture density $H()$ is a function of a multivariate D-dimensioned response Y conditioned on the predictor variables X, concomitant variables w and model parameters $\psi = \{\alpha, \theta\}$. The concomitant parameters α define the probability π of each observation belonging in latent class k based on the values of w. The membership function π is a discrete response model, a multinomial logit model in our case. The relationship between the predictor variables X and each dependent variable Y_d is defined by the parameters θ_{kd} of the mixture function f_{kd}, which in our case is a Gaussian linear regression. For this initial analysis we restrict the number of classes to two, $k = 2$. We estimate the model using an expectation-maximization algorithm included in the `flexmix` package for R Leisch (2004).

This analysis has three goals: (1) identify whether the Fulton county housing market should be considered as more than one distinct market based on multidimensional value performance, (2) determine the variables that lead to inclusion in a more "favorable" market, and (3) determine if the markets show different relationships between the covariates and their outcomes.

We estimated four models, though we only present the model with likelihood sufficient to reject the other three. The first is an intercepts-only model with $\pi_k = \Lambda(\boldsymbol{\iota}\alpha_k)$ and $f_{kd} = \boldsymbol{\iota}\theta_{kd}$. The second is a concomitants-only class membership model using the set of covariates from Section 4.3 (including the interaction effect of income growth and transit proximity); thus $\pi_k = \Lambda(X\alpha_k)$ and $f_{kd} = \boldsymbol{\iota}\theta_{kd}$. The third model is the complement to the second, with the set of covariates used in the response functions; $\pi_k = \Lambda(\boldsymbol{\iota}\alpha_k)$ and $f_{kd} = X\theta_{kd}$. The fourth model includes a full set of covariates in the class membership function as well as the response functions, $\pi_k = \Lambda(X\alpha_k)$ and $f_{kd} = X\theta_{kd}$.

The model results allow us to reject the null hypothesis that $k = 1$, indicating that there are at least two latent classes of home price performance in the Atlanta region. The model predicts that most homes belong to Class 1 (91.7%). Figure 7 shows the price performance for a random sample of homes drawn from the model's membership assignments. The difference in trends between the two classes is immediately apparent. Homes in Class 1 experienced modest growth in prices until 2009, and then collapsed in value to a varying degree. Homes in Class 2, conversely, gained a substantial amount of value in the middle part of the period, and the average in the class remains higher than its initial value. Homes in Class 2 are likely to have a higher average value, a higher average growth rate, but also greater variance in the growth rate. Based on the analysis in the previous section, this cursory observation leads us to expect homes in Class 2 will be closer to MARTA stations.

Indeed, it appears that this is the case. Figure 8 compares the spatial density distribution for homes in Class 1, Class 2, and the full dataset. Class 1 is distributed throughout the county, but Class 2 is concentrated near the city center. The effect of proximity to MARTA on membership in Class 2 is further confirmed by the estimated parameters for the concomitant model given in Table 16. In this model positive estimates indicate a home is more likely to belong in Class 2 than Class 1 as the associated concomitant variable increases. With the exception of the distance to a freeway entrance, each of the estimates is highly significant. Membership in Class 2 is more likely as the property size increases, the number of rooms in the home decreases, the age of a home decreases, the median neighborhood income decreases, and as the percent of white homes increases. Also, and

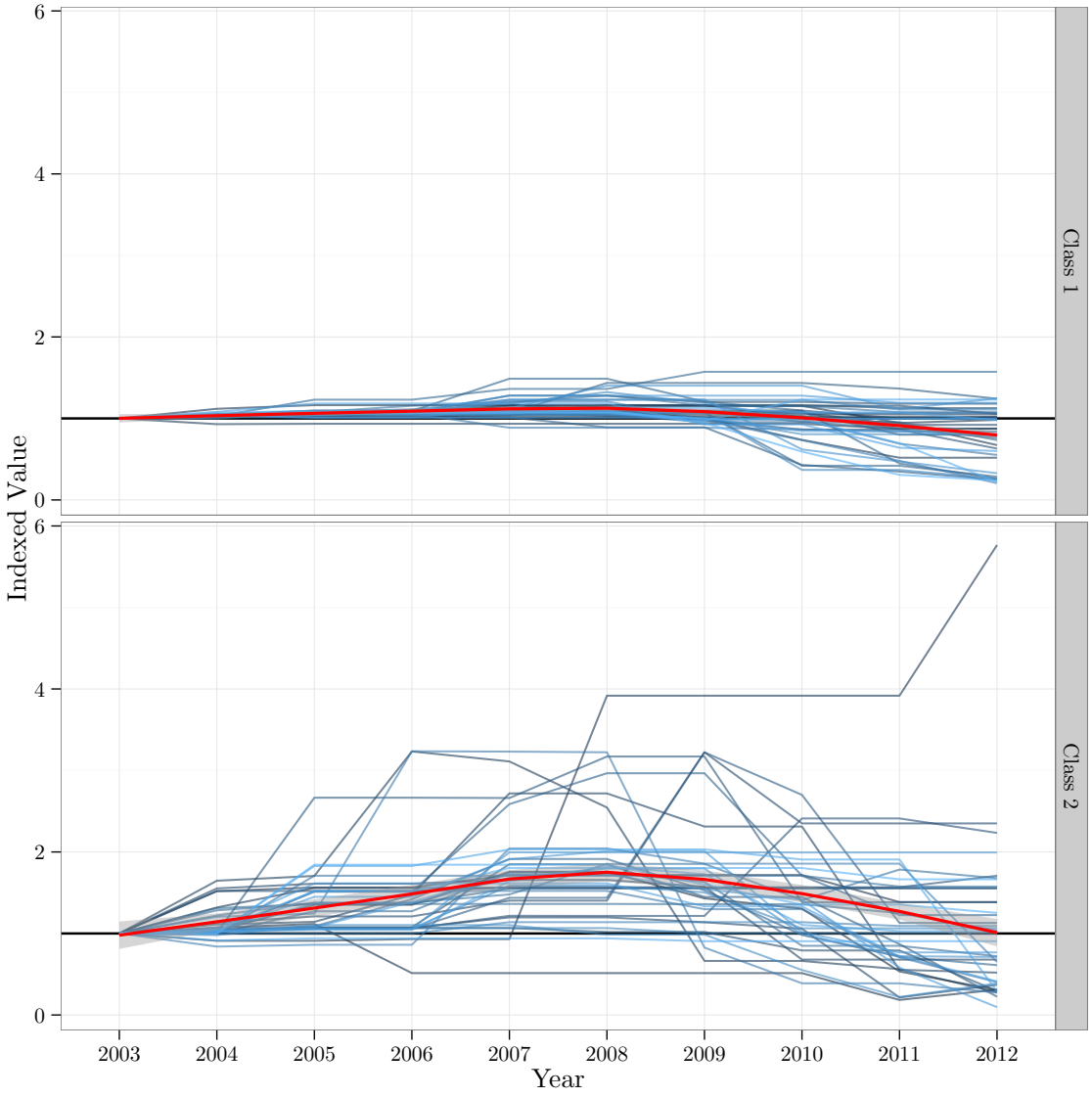

Figure 7: Value performance for a sample of 40 homes in each latent class.

Table 16: Concomitant (class membership) model.

	Class 2	
Covariate	α	t-stat
(Intercept)	23.0^{***}	9.69
log(Acres)	0.522^{***}	6.14
Total Rooms	-0.204^{***}	-4.75
log(Effective Age)	-0.702^{***}	-6.74
log(Income)	-1.21^{***}	-5.47
Percent White	0.931^{**}	2.68
log(Distance to Freeway)	-0.087	-0.823
Income Growth	-7.03^{***}	-4.3
log(Distance to Rail)	-0.801^{***}	-8.64
log(Rail Dist.)$*$ Income Growth	0.864^{***}	4.72

** significant at $p < .01$; $^{***}p < .001$

more relevant to our study, membership in Class 2 is less likely as the distance between a home and the nearest MARTA station increases.

The effect of neighborhood income growth and its interaction with transit proximity is more complicated than in the univariate SDM analysis. The negative coefficient estimates of the income growth and distance to rail parameters indicate that as neighborhood income growth or the distance to rail increases, homes are less likely to be in Class 2. But the significant and positive interaction term implies that at higher levels of income growth, homes further from transit are *more* likely to be in Class 2.

An important feature of latent class mixture models is that the parameters θ_{kd} are allowed to have different effects on the response for each class. Figure 9 presents the coefficient estimates for the response models visually to aid comparison.[5] As the confidence intervals for many estimates overlap, we cannot reject that the effects of most covariates are the same across classes. There are some notable exceptions, however. The age of a home has no effect on any of the response variables for homes in Class 1, but a significant negative effect on each of the response variables for homes in Class 2. Simply, older homes

[5]In this figure, the estimates and confidence intervals on the "Income Growth" variable have been scaled by 0.10 to improve clarity.

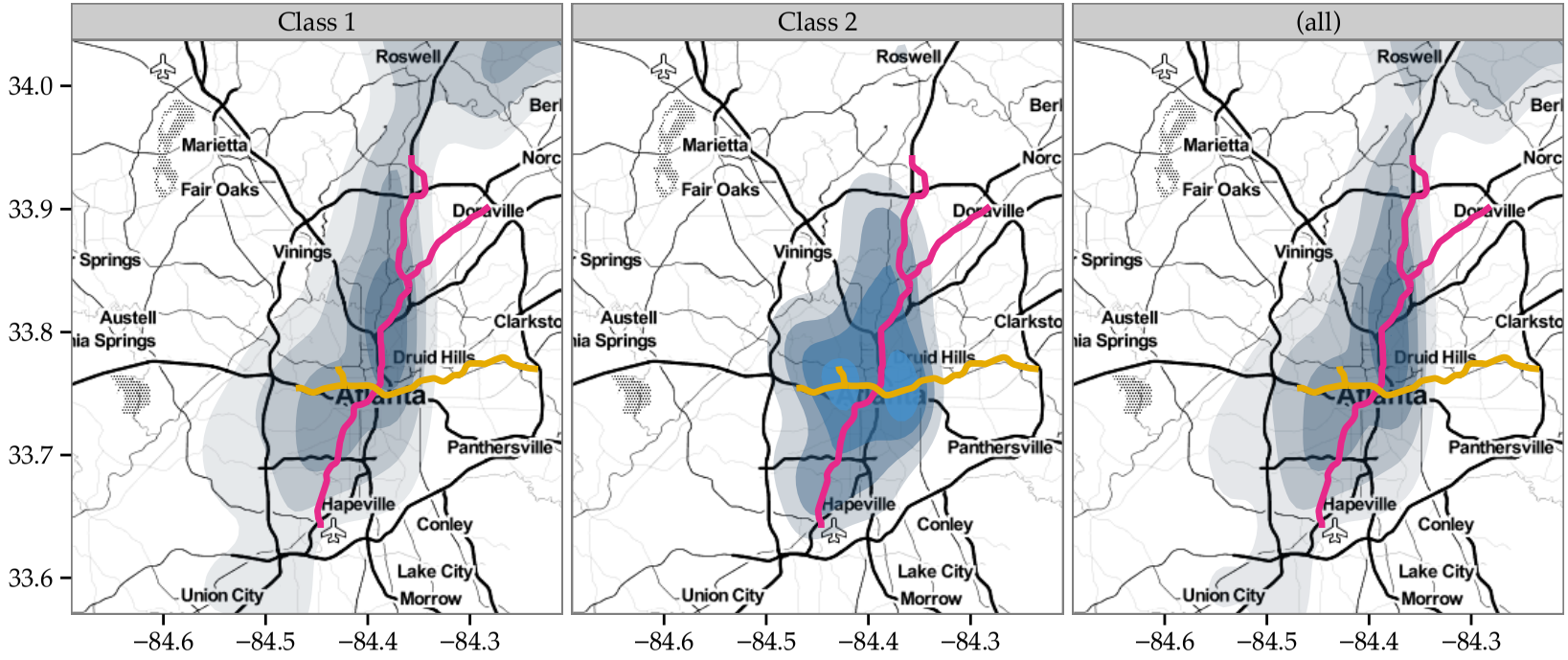

Figure 8: Relative spatial density of homes in each latent class.

are less valuable and experience lower growth only if the home is in Class 2, though older homes also have a less volatile growth rate. Conversely, the percent of white people in a neighborhood is more significantly correlated with the response variables for homes in Class 1. For these homes, their value and growth rate increase with the share of white neighbors, and variance in the growth rate decreases. The effects of income growth also differ for each class, increasing value volatility in Class 2 and decreasing it in Class 1.

In terms of the correlation between transit infrastructure and the response variables, the evidence is mixed. For homes in Class 1, proximity to MARTA rail is significantly correlated with higher average values, higher average growth rates, and lower growth volatility. But none of these correlations is significant for homes in Class 2. The same is true for freeway access, though the effects are smaller. But the interaction term with income growth is significant in all the response models, suggesting that in gentrifying neighborhoods closer to transit, mean values and mean growth rates are higher. But for volatility, the signs are reversed again.

4.5 Interpretations and Future Directions

The empirical results in general sustain the theories presented in Section 4.1.1. As the model of Glaeser et al. (2008) predicted, homes near MARTA have higher expected values but also higher volatility. And as the model of Guerrieri et al. (2013) predicted, homes in neighborhoods with increasing average incomes showed higher growth rates than homes in other neighborhoods; there was some additional effect for these neighborhoods if they were closer to transit stations. The significance of the interaction between income growth and distance to MARTA highlights that income growth in a Census tract can occur by two very different means: it could be a result of typical gentrification processes, with wealthy people displacing the poor and/or redeveloping former industrial zones; but income growth in a tract can also result from urbanization or suburban sprawl, as people move on to previous undeveloped land. Our analysis results suggest that it is the first process that leads to higher home values.

The two economic theories are complementary in that the Glaeser et al. model does

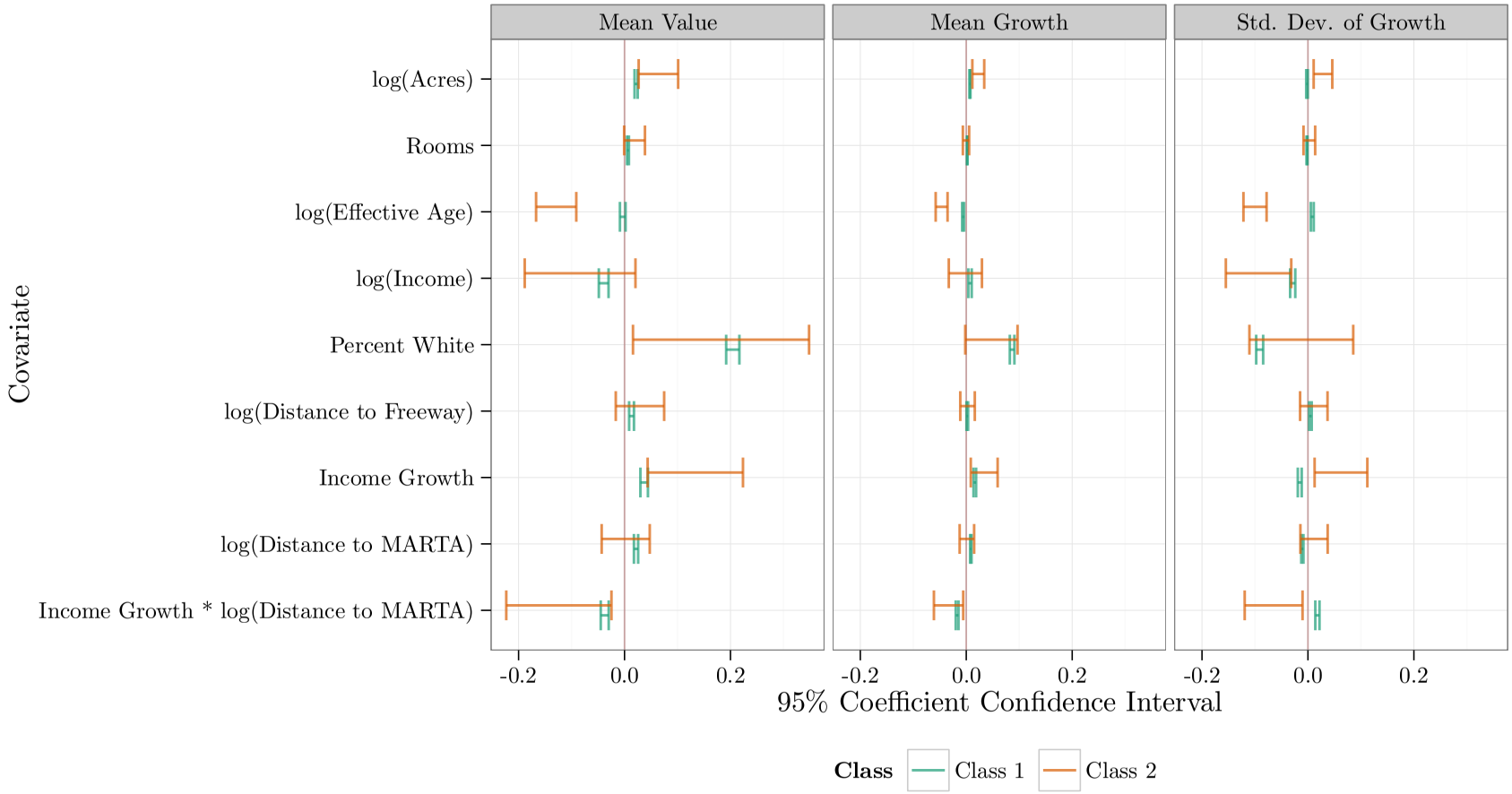

Figure 9: Latent class covariate estimates and confidence intervals.

not explain why markets may be inelastic, and the Guerrieri et al. model is agnostic to the cause of the demand shock. But the Guerrieri et al. model does not allow that the demand shock is irrational or even temporary, and thus provides no prediction on *where* prices will decline. Our empirical results, particularly the latent class membership analysis in Figure 7 and Table 16, suggests that prices increase primarily in the city center but decrease globally. Though the model coefficients suggest that transit proximity plays a role in this, we cannot rule out that there may be missing or endogenous variables for which MARTA is merely an indicator, such as population density or a loosely defined "urbanness."

The univariate analysis presented in Section 4.3 controls these missing variables and shows many of the same results, but without the subtleties of the multivariate, latent class analysis in Section 4.4. As an illustration of the differences between the models, consider the effects of income growth and rail proximity on growth rate volatility. In the spatial models none of the three coefficients is significant, but in the latent class model four of the six coefficients are significant. It is impossible to know how much of this discrepancy results from spatial dependence or correlation in the latent class models, and how much results from aggregation bias in the spatial models that assign a single coefficient estimate to at least two separate categories.

A potential resolution to this discrepancy is to introduce spatial effects into the mixture model, and these effects may need to be introduced in both the membership model as well as the response models. Wall and Liu (2009) present a univariate latent class model that introduces spatially correlated errors into the membership model. Gelfand and Vounatsou (2003) present a (single class) multivariate mixture model for continuous response variables that incorporates spatial dependence. We have not seen spatial effects introduced at both levels in our initial explorations of the literature. Developing such a model would be an important methodological contribution, as well as a necessary step in fully understanding this problem.

4.6 References

Alonso, W., 1960. A theory of the urban land market. *Papers in Regional Science* 6, 149–157.

Anderson, C.W., Beracha, E., 2010. Home price sensitivity to capital market factors: Analysis of ZIP code data. *The Journal of Real Estate Research* 32, 161–185.

Bivand, R., 2013. spdep: Spatial Dependence, Weighting Schemes, Statistics, and Models. URL: `http://cran.r-project.org/package=spdep`.

Bowes, D.R., Ihlanfeldt, K.R., 2001. Identifying the impacts of rail transit stations on residential property values. *Journal of Urban Economics* 50, 1–25.

Debrezion, G., Pels, E., Rietveld, P., 2007. The impact of railway stations on residential and commercial property value: a meta-analysis. *The Journal of Real Estate Finance and Economics* 35, 161–180.

Gelfand, A.E., Vounatsou, P., 2003. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 4, 11–25.

Glaeser, E.L., Gyourko, J., Saiz, A., 2008. Housing supply and housing bubbles. *Journal of Urban Economics* 64, 198–217.

Grass, R.G., 1992. The estimation of residential property values around transit station sites in Washington, D.C. *Journal of Economics & Finance* 16, 139–146.

Guerrieri, V., Hartley, D., Hurst, E., 2013. Endogenous gentrification and housing price dynamics. *Journal of Public Economics* 100, 45–60.

Iacono, M., Levinson, D., 2011. Location, regional accessibility, and price effects: evidence from home sales in Hennepin County, Minnesota. *Transportation Research Record* 2245, 87–94.

Leisch, F., 2004. FlexMix : A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* 11, 1–18.

Lewis-Workman, S., Brod, D., 1997. Measuring the neighborhood benefits of rail transit accessibility. *Transportation Research Record* 1576, 147–153.

McMillen, D.P., McDonald, J., 2004. Reaction of house prices to a new rapid transit line: Chicago's Midway Line, 1983-1999. *Real Estate Economics* 32, 463–486.

Pivo, G., 2013. The effect of transportation, location, and affordability related sustainability features on mortgage default prediction and risk in multifamily rental housing. URL: `http://www.fanniemae.com/resources/file/aboutus/pdf/hoytpivo_mfhousing_sustainability.pdf`.

Pollack, S., Bluestone, B., Billingham, C., 2010. *Maintaining diversity in Americas transit-rich neighborhoods: Tools for equitable neighborhood change*. Technical Report 3. Dukakis Center for Urban and Regional Policy. URL: `http://iris.lib.neu.edu/dukakis_pubs/3/`.

Smith, J.J., Gihring, T.A., 2006. Financing transit systems through value capture. *American Journal of Economics and Sociology* 65, 751–786.

Standard & Poors, 2013. Case-Shiller Home Price Index. URL: `http://us.spindices.com/index-family/real-estate/sp-case-shiller`.

U.S. Census Bureau, 2010a. *B02001. Race: 2010 ACS 5-year estimates*. Technical Report.

U.S. Census Bureau, 2010b. *B19013. Median household income in the past 12 months: 2010 ACS 5-year estimates.* Technical Report. URL: `http://factfinder2.census.gov/`.

Wall, M.M., Liu, X., 2009. Spatial Latent Class Analysis Model for Spatially Distributed Multivariate Binary Data. *Computational statistics & data analysis* 53, 3057–3069.

CHAPTER V

CONCLUSION

5.1 Summary of Findings

As stated in the introductory chapter, this dissertation has two primary objectives. The first objective is to demonstrate that emerging "big" data sources can be used to estimate different types of travel behavior or land use models with results comparable to more traditional methods. The second objective is to illustrate examples of additional questions that can be explored with such data, taking advantage of the new variables and large sample sizes available with these data.

The vehicle ownership model presented in Chapter 2 shows the same effect of various household socioeconomic characteristics and built environment attributes on vehicle ownership rates as have been found in the previous literature. Based on this observation, the pairing of targeted marketing records and vehicle registration databases seems a promising strategy for estimating vehicle ownership models used in regional travel models and further research projects. Similarly, the land value model estimated in Chapter 3 produced findings consonant with both intuition and the relevant literature. Though the county assessor records used in Chapter 4 have long been used to examine the relationship between transportation infrastructure and the housing market, in this case they serve to confirm the spatial analyses of Chapter 3.

In terms of the new analyses that big data records make possible, Chapter 2 examines a question that could not be easily addressed by other means. In a way, the primary purpose of personal credit records is to record a person's history, and therefore are particularly suited to examining questions related to past experience. Though the studies presented in Chapters 3 and 4 did not use variables that are unusual or rare in such studies, characteristics of the datasets eased or enabled the analysis in different ways. The disaggregate nature of the targeted marketing data in Chapter 3 permitted an analysis of the direct and indirect

effects of each covariate on a home's value. Most similar studies (including that in Chapter 4) use socioeconomic information that is aggregated to the home's neighborhood at some scale; the similarity between the results in Chapters 3 and 4 suggests that in this case, the aggregation of socioeconomic variables at the neighborhood level is not a concern. What these datasets do show, however, is the advantage that can come from large sample sizes. Even though computational limitations prevented the analysis in Chapter 4 from running on the full dataset, the analysis sample remains large relative to previous work, and the study is able to identify significant relationships that may not have been identified with fewer observations.

5.2 Directions for Future Research

As an initial investigation into these types of data, there are many remaining questions and associated opportunities for further investigation. These include the estimation of additional models, expanding the analysis to other regions and times, and applying big data methodologies.

This dissertation presented examples of only two models of the several identified in the introductory chapter as being important for regional policy analysis. Most importantly, none of the four models comprising the common "four-step" travel modeling system were presented. Though the targeted marketing records and public administrative databases used throughout this dissertation contain a great wealth of socioeconomic and related variables, they do not contain information on the households' actual trip-making behavior. The motivation for the study in Chapter 2 assumes that households with more vehicles make more vehicle trips; though this has been suggested in the literature (Giuliano and Dargay, 2006), vehicle ownership does not by itself provide information about the number, types, and destinations of the owner's trips. Similarly, households who are willing to pay to live close to transit may use transit for all of their trips, or only certain types to a particular set of destinations. Developing a methodology to create databases that would allow these types of models is an essential next step in ascertaining the usefulness of these data sources.

One of the potential advantages of big data resources is the ability to join many different

types of database together. As an example, a great deal of the additional information required to estimate other travel behavior models may come from the electronic device traces described in Chapter 1, which may provide the positions of people through the day and the routes they took to get to each point. Related to this is the idea that the databases can be joined to themselves, but at different time points. The panel dataset suggested as ideal to answer the remaining questions regarding the causal effects of prior experience in Chapter 2 could be created by joining the dataset used in that analysis to an identical dataset taken a few years later or earlier.

Another potential advantage of big data resources is their uniformity across regions. Each of the studies in Chapters 2 through 4 used data specific to the Atlanta metropolitan area, but the methodologies should be generally applicable. Replication of results is an increasingly important part of scientific research (Hamermesh, 2007), and the transferability of modeling strategies and innovations between cities is an important consideration for travel demand modeling practice. Big datasets promise to make this easy, as the same analysis code could operate on a dataset from the same provider but with a different scope or focus. Determining the actual feasibility of this should be an important research objective.

A final recommendation for future research returns to the definition of big data. As Gartner (2013) defines it, big data is

> high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

The data used in this dissertation certainly qualifies under this definition, but actual methodologies are those that might just as easily be used on "small" data. Analyses employing the "innovative forms of information processing" such as active database queries and non-parametric data mining techniques would be an important next step. Though active queries may not be available to researchers that do not have direct access to the databases (these may only be available from inside the data providers), a sufficiently large static dataset should still provide opportunities for data mining, or identifying patterns and

correlations within the data that may not have been identified through standard methods of hypothesis testing (Yu, 1996; Hegland, 2001). The latent class mixture model presented in Chapter 4 could be considered a preliminary step in this direction.

If the history of transportation modeling data is any guide, the question of how big data resources might be used in transportation planning and modeling will remain important to the field for the foreseeable future.[1] Identifying which types of database are useful for which types of analysis will likewise remain an important methodological skill for transportation researchers and practitioners. This dissertation, though only an initial investigation, illustrates the potential of these databases, in particular the utility and flexibility of targeted marketing records. The disaggregate socioeconomic information available on these records is a crucial component of many transportation studies, and one that promises to become ever more important as other source databases mature.

5.3 References

Gartner, 2013. Big Data. URL: `http://www.gartner.com/it-glossary/big-data/`. accessed November 25, 2013.

Giuliano, G., Dargay, J., 2006. Car ownership, travel and land use: a comparison of the US and Great Britain. *Transportation Research Part A: Policy and Practice* 40, 106–124.

Hamermesh, D.S., 2007. Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue canadienne d'économique* 40, 715–733.

Hegland, M., 2001. Data mining techniques. *Acta Numerica* 10, 313–355.

Yu, P., 1996. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8, 866–883.

[1] *Transportation* and *Transportation Research Part C* currently have open calls for papers addressing "emerging, passively collected datasets for examining travel behavior" and "big data in transportation and traffic engineering," respectively.

APPENDIX A

APPENDIX TO THE VEHICLE OWNERSHIP PAPER

We tested the sensitivity of model results to several data processing and modeling assumptions. Since we know that the number of children variable is incomplete for some households that do in fact have children, we applied various imputation methods to correct for this issue; however none of the imputation methods resulted in different model interpretations.

We used the duration exposure metric to test the sensitivity of results to our "previous move" assumptions, described in Section 2.3.3. Specifically, we calculated the duration exposure metric for the earliest possible move-in date and latest possible move-in dates using $\alpha = 1$, which applies an equal weight to each prior exposure day. The impact on the exposure metric was minimal: only 497 of the 227,830 records changed by more than 1%. Since results were robust to both of our previous move assumptions, we only report those models that use the latest move assumption in Table 4.

The MNL model imposes the "independence of irrelevant alternatives" (IIA) assumption, requiring that the unobserved characteristics of each alternative j (ϵ_{ij}) are independent of both the unobserved characteristics ($\epsilon_{ij'}$) and the observed utility ($V_{ij'}$) for all other alternatives j'. This assumption may be violated for ordinally-related alternatives, since it is plausible that similar unobserved characteristics would influence the choice between adjacent alternatives in particular. To test whether the MNL model was appropriate, we applied a Hausman-McFadden IIA test (Hausman and McFadden, 1984) comparing the estimates of each exposure model to models with an identical specification but an alternative removed. The tests failed by producing a negative statistic; Small and Hsiao (1985, p. 619) point out that such computational failures are not uncommon with the Hausman-McFadden test, given that it "requires inversion of the difference between two closely related matrices [the variance-covariance matrices of the coefficient estimators for the reduced-choice-set and full-choice-set models], which may be non-positive-definite or nearly singular." We also

tested a nested specification of the extreme model that allowed for correlated substitution between the two- and three-or-more-vehicle alternatives. The estimated nest substitution parameters were all greater than 1, implying a violation of random utility theory and the other coefficients were not materially different from the extreme exposure MNL model. We therefore retain the MNL specification.

A.1 References

Hausman, J., McFadden, D., 1984. Specification tests for the multinomial logit model. *Econometrica* 52, 1219–1240.

Small, K.A., Hsiao, C., 1985. Multinomial logit specification tests. *International Economic Review* 26, 619–627.

APPENDIX B

APPENDICES TO THE SPATIAL AUTOREGRESSION PAPER

B.1 Sampling Bias

The sampling frame for the study data is the Georgia Motor Vehicles registration database. This raises the possibility that the sample is unrepresentative of households in the Atlanta region, as households owning multiple vehicles have a higher likelihood of being sampled, and households that do not own vehicles — or that only lease vehicles — are excluded. This is only a problem, however, if zero-vehicle households are common among home owners, as we excluded renters from our analysis. To examine the potential for unrepresentativeness in our sample, we compared our analysis data with the 2006-2011 5-year aggregated public use microsample (PUMS) file representing Fulton and DeKalb counties from the American Community Survey (ACS).

According to the PUMS data, 587 households of the 30,381 respondents in Fulton and DeKalb counties owned a home but did not own a vehicle, implying that we failed to sample approximately 3% of the relevant households. Our data contain a different set of variables than the ACS questionnaire, and we therefore cannot compare the datasets variable-to-variable. For the household income variable, however, we were able to run a Kolmogorov-Smirnov test comparing the distribution in our sample versus that in the ACS PUMS: we rejected that the two distributions were the same with a p-value of 0. Figure 10 illustrates where our sample differs from the ACS microdata: we observe fewer households with incomes over \$250k, but more in the \$100k to \$175k range. The results of this comparison analysis suggest that our sample may not be perfectly representative of the Atlanta housing market, but not likely enough to seriously compromise our findings.

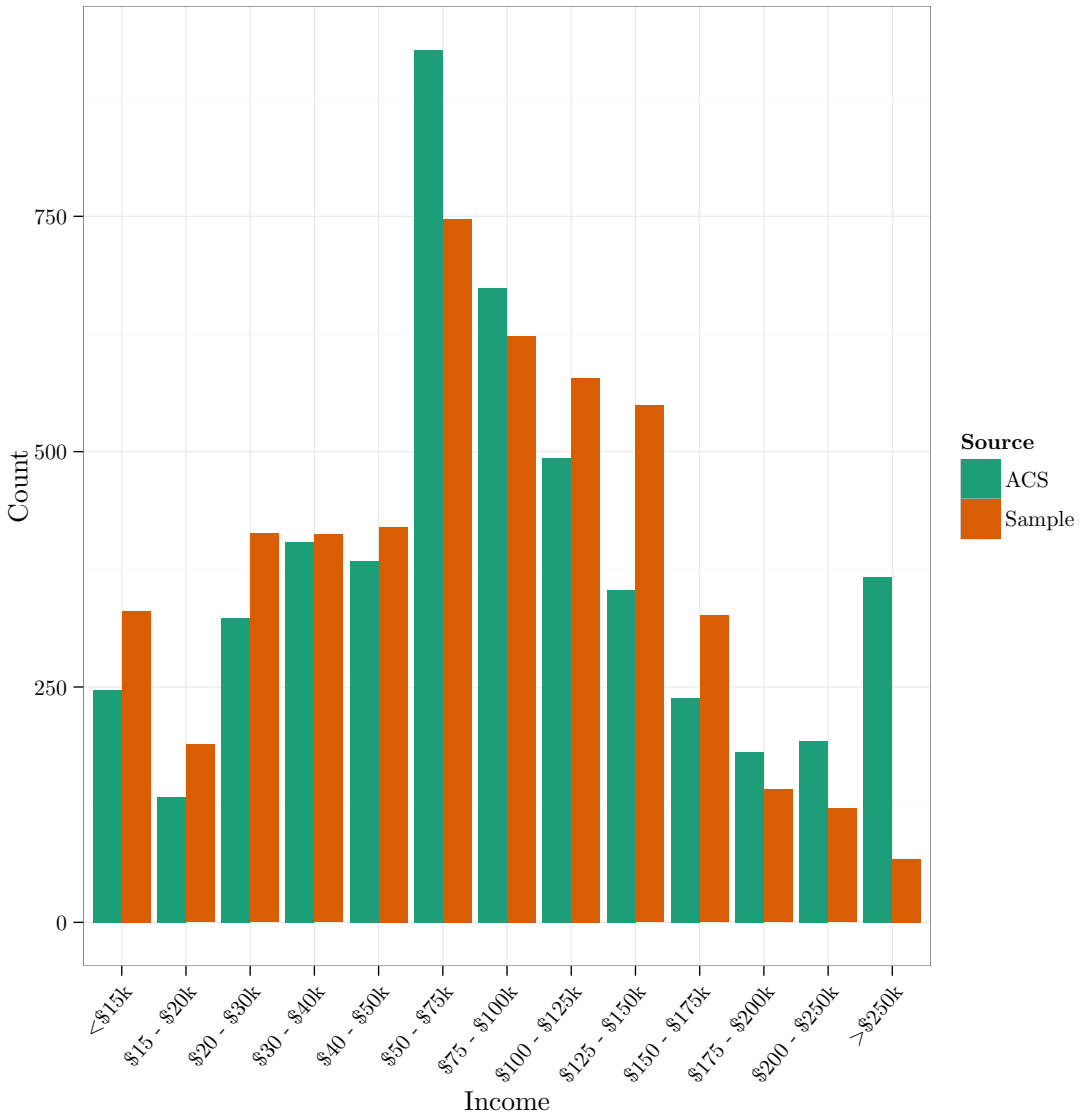

Figure 10: Distributions of incomes in the ACS and our estimation sample.

B.2 Spatial Weights

There are a number of ways to specify spatial weights matrices. The most common spatial simultaneous autoregressive models were originally developed for areal regions, and this has led to a number of specifications that are not primarily intuitive. There are three basic species of weights matrix:

1. Adjacency: do Voronoi polygons around the observations touch? This can be extended to higher orders.

2. Nearest k observations, regardless of distance.

3. Observations within distance d, regardless of number.

The nearest observations and distance radius methods may both be weighted by distance to assign higher value to nearer observations. This creates five candidate schema, each with an array of inclusion possibilities (by allowing k or d to increase).

In this appendix, we examine the model likelihood and parameter stability of a spatial Durbin model using each of the five candidate schema with nine different inclusion rules. For the Voronoi polygons, we considered 1^{st} through 9^{th}-order adjacency. For the k-nearest neighbors method, we use $k = 2$, 5, 10, 15, 20, 35, 50, 75, and 100. For the radius method, we use nine equal divisions of the the range $d = [0.5, 4.0]$ miles. We also consider an inverse distance weighting scheme for both the k-nearest and d radius schema.

Figure 11 shows the log-likelihood of our SDM specification estimated for each of the candidate weights matrices. As shown in the chart, the Voronoi polygon method produces its maximum log-likelihood for first-order contiguity, and drops substantially as higher orders are considered. The k-nearest neighbors method produces its maximum at 20 neighbors, a much higher level than used by either Löchl and Axhausen (2010) or Ibeas et al. (2012). Considering neighbors within a particular radius has its highest likelihood at 1.4 miles. Weighting both the k-nearest and d-radius methods substantially improves the maximum achieved likelihood for both methods, with the optimum number of neighbors being 50 and the optimum radius now 1.8 miles.

LeSage and Pace (2009) assert that the particular weighting scheme should not have a serious influence on the estimated parameters. Our findings presented in Figure 12 provide some initial support for that claim, but also some disputations. The autocorrelation parameter ρ increases monotonically with expanding inclusion, with the notable exception of the Voronoi polygon method, which drops drastically above 7^{th}-order adjacency. The direct coefficient β and the indirect coefficient γ are usually opposites; for instance, the weighted k method has the most positive β but also the most negative γ, potentially muting its effect. We selected the weighted radius method because it has a high model likelihood and conservative coefficent estimates, falling as they do in the middle of the range defined by the candidate weighting scheme. Establishing which scheme best represents a particular housing market is an important opportunity for further research, and there may not be a general answer.

B.3 References

Ibeas, A., Cordera, R., Dell'Olio, L., Coppola, P., Dominguez, A., 2012. Modelling transport and real-estate values interactions in urban systems. *Journal of Transport Geography* 24, 370–382.

LeSage, J.P., Pace, R.K., 2009. *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.

Löchl, M., Axhausen, K.W., 2010. Modeling hedonic residential rents for land use and transport simulation while considering spatial effects. *Journal of Transport and Land Use* 3, 39–63.

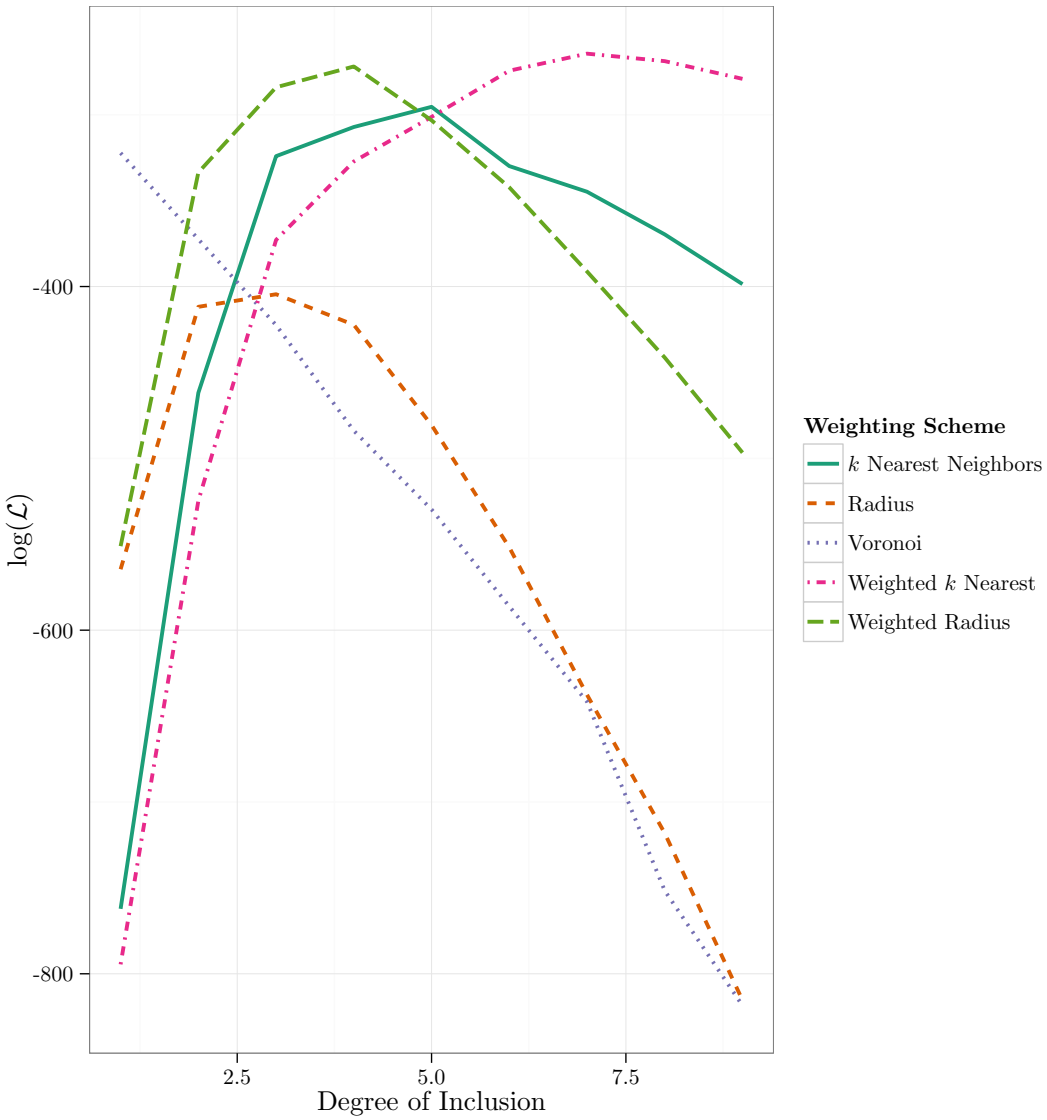

Figure 11: Log-likelihood under different weighting regimens.

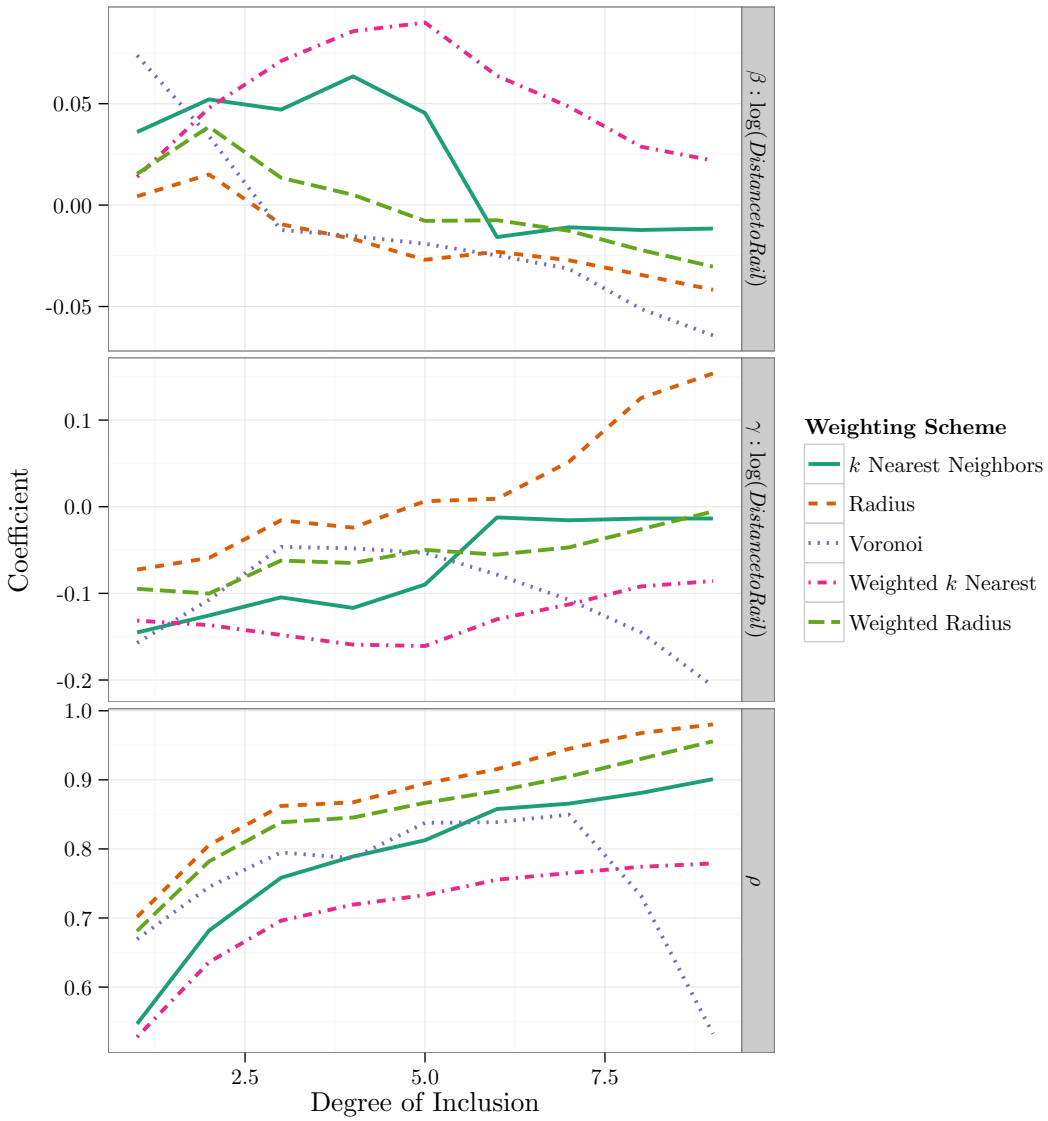

Figure 12: Autocorrelation and distance to rail parameters under differing weighting regimens.

VITA

Gregory S. Macfarlane is a native of Provo, Utah, and an alumnus of Brigham Young University (Civil Engineering '09) and Provo High School. His interest in transportation studies originates from his experiences as a Latter-day Saint missionary in Singapore, Malaysia, and Sri Lanka.